Analysis of Complex Survival and Longitudinal Data in Observational Studies

by

Fan Wu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2017

Doctoral Committee:

Professor Yi Li, Co-Chair
Research Assistant Professor Sehee Kim, Co-Chair
Assistant Professor Sung Kyun Park
Professor Emeritus Roger Wiggins
Associate Professor Min Zhang

Truly, truly, I say to you,
unless a grain of wheat
falls into the earth and dies,
it remains alone;
but if it dies,
it bears much fruit.
—John 12:24

To Chen

ACKNOWLEDGEMENTS

I would like to thank my Advisors, Dr. Yi Li and Dr. Sehee Kim, whose support and guidance have helped me during the past five years in both my research and my life. Yi has funded my since I entered the program as a doctoral student. He has given me much freedom in choosing the topics for my research, and provided me with his instruction and inspiration whenever I meet obstacles. I am deeply grateful to Sehee for all her effort and time spent on revising my manuscripts. This work would not have been completed without the back-to-back meetings with her.

Special thanks go to my other committee members. Dr. Min Zhang has been giving me constructive suggestions since I took her repeated measures class. It is a great pleasure that I had the opportunity to work with her on the third project. I would like to thank Dr. Sung Kyun Park for providing the Normative Aging Study data, and giving useful comments from the point of view of an experienced epidemiologist. My sincere gratitude goes to Dr. Roger Wiggins, whose passion about research has been a real inspiration for me. I have learned so much from his expertise in kidney diseases.

Thanks are due to Dr. Dorota Dabrowska from the University of California, Los Angeles. She has been very supportive during my application for doctoral study, which gave me the chance to join Michigan in the first place. With her rich knowledge in survival analysis, she provided me a lot of advices for my projects on the left-truncated data.

I would like to thank my friends and colleagues at the University of Michigan. During my difficult time trying to figure out the asymptotic proofs, the study group with Kevin, Yanming and Fei gave me the very first introduction to empirical processes. Wenting and Zihuai have alway been there and ready to lend me a hand at times when I need help.

No words can express my gratitude for the full and hearty support of my parents for my study and research. Though they may not understand my work, their unconditional love has always soothed and comforted me over the years. Lastly, I would like to thank Chen and Sasa. It could have taken me less time to finish this dissertation without them giving me so much joyful memories, or it could have never been finished at all.

TABLE OF CONTENTS

DEDICATION . . . . . **ii**

ACKNOWLEDGEMENTS . . . . . **iii**

LIST OF FIGURES . . . . . **vii**

LIST OF TABLES . . . . . **viii**

LIST OF APPENDICES . . . . . **ix**

ABSTRACT . . . . . **x**

CHAPTER

I. Introduction 1

II. Literature Review 5

2.1 Length-Biased Sampling Methods 5
2.2 Composite Likelihood Methods 11
2.3 Clustering Methods for Longitudinal Data 15

III. A Pairwise Likelihood Augmented Cox Estimator for Left-Truncated Data 19

3.1 Introduction 19
3.2 Proposed Method 21
3.2.1 Preliminaries 21
3.2.2 Pairwise-Likelihood Augmented Cox (PLAC) Estimator 23
3.2.3 Asymptotic Properties 27
3.3 Simulation 31
3.4 Data Application 35
3.5 Discussion 37

IV. A Pairwise Likelihood Augmented Cox Estimator with Application to the Kidney Transplantation Registry of Patients under Time-Dependent Treatments 41

4.1 Introduction 41
4.2 Proposed Method 44
4.2.1 Preliminaries 44
4.2.2 The PLAC Estimator for Data with Time-Dependent Covariates . 46
4.2.3 The Modified Pairwise Likelihood 49
4.3 Simulation 51

4.4 Data Application 55
4.5 Discussion 61

V. Longitudinal Data Clustering Using Penalized Least Squares 64

5.1 Introduction 64
5.2 Proposed Method 68
5.2.1 Clustering Using Penalized Least Squares 68
5.2.2 Cluster Assignment 70
5.2.3 Comparing Clusterings 72
5.3 Simulation 73
5.4 Data Application 77
5.5 Discussion 80

VI. Conclusions and Future Work 83

APPENDICES **86**

BIBLIOGRAPHY **131**

LIST OF FIGURES

Figure

3.1 Estimated Survival for Patients with or without Diabetes in the RRI-CKD data. . 38
4.1 Examples of different follow-up scenarios in left-truncated right-censored data. . . . 50
4.2 Christmas tree plot for the coefficient estiamtes for PD and TX in the UNOS data. 58
4.3 US maps of hazards ratio estimates for PD and TX compared with HD. 59
5.1 Illustration of the clustering gain. 71
5.2 Clustering results for SBP. 78
5.3 Clustering results for DBP. 79
A.1 Estimated survival curves of A and V for RRI-CKD data. 105
A.2 Estimated hazards ratios of the the covariates in the RRI-CKD data. 106
A.3 Estimated $\hat{G}$ for each level of the covariates in the RRI-CKD data. 107
C.1 The profiles of the true cluster centers used in the simulation. 128
C.2 Example trajectories for Simulation I, Case 1 129
C.3 Example trajectories for Simulation I, Case 2 129
C.4 Example trajectories for Simulation II 130

LIST OF TABLES

Table

3.1 Summary of simulation with various sample sizes and censoring rates. 33
3.2 Coefficient estimates from the RRI-CKD data. 37
4.1 Summary of simulation with various cases for $Z_v(t)$. 53
4.2 Summary of simulation with various G under Case 1 with no censoring. 55
4.3 Coefficient estimates for UNOS transplantation data in OH and WV. 60
5.1 Mean clustering index under different within-cluster heterogeneity, measurement errors, and coefficient distributions. 75
5.2 Mean clustering index under various sparsity of the observations. 76
5.3 Cross table of cluster memberships for SBP and DBP. 79
5.4 Demographics, smoking history, and hypertension (HT) comparison for the SBP and DBP clusterings. 80
A.1 Summary of simulation with $N = 200$ and various censoring rates. 103
A.2 Summary of simulation using transformation approach. 104
B.2 Summary of simulation in Case 2 with various G. 120
B.1 Summary of simulations with various sample sizes. 121
B.3 Summary of simulation in Case 3 with various G. 122
B.4 Summary of simulation in Case 1 with various F_ζ. 123
B.5 Sample sizes and censoring rates for the UNOS datasets. 124

LIST OF APPENDICES

Appendix

A. Proofs, Additional Simulation and Data Analysis for the First Project 87

A.1 Proofs of the Asymptotic Properties for the Pairwise Likelihood Augmented Cox Estimator 87
A.1.1 Identifiability 89
A.1.2 Consistency 91
A.1.3 Asymptotic Normality 96
A.2 Additional Simulation Results 103
A.3 Additional Data Analysis Results 105

B. Proofs, Additional Simulation and Data Analysis for the Second Project 108

B.1 Asymptotic Properties of the PLAC Estimator for Time-Dependent Covariates 108
B.2 Additional Simulation Results 120
B.3 Additional Data Analysis Results 124

C. Algorithm, Simulation Setup and Data Analysis Results for the Third Project 125

C.1 An Alternating Direction Method of Multiplier 125
C.2 Simulation Setups 128

ABSTRACT

Analysis of Complex Survival and Longitudinal Data in Observational Studies

by

Fan Wu

Co-Chairs: Yi Li, PhD and Sehee Kim, PhD

This dissertation is motivated by several complex biomedical studies, where challenges arise from that 1) survival data from a prevalent cohort are subject to both left truncation and right censoring, and 2) longitudinal data on human subjects are sparse and unbalanced. For example, in the Renal Research Institute Chronic Kidney Disease (RRI-CKD) study and in the United Network for Organ Sharing (UNOS) kidney transplantation registry, recruited were patients with kidney diseases of which the onsets precede the enrollment, whereas in the Normative Aging Study (NAS), subjects' measurements were not collected at a common sequence of ages. There is an urgent necessity to develop robust and efficient methods to analyze such data which account for their observational nature. This dissertation, comprising of three projects, proposes a cohort of new statistical methods to address these challenges.

In the first project, we consider efficiency improvement in the regression method with left-truncated survival data. When assumptions can be made on the truncation, conventional conditional approaches are inefficient, whereas methods assuming parametric truncation distributions are pruned to misspecification. We propose a pairwise likelihood augmented Cox estimator assuming only independence between

the underlying truncation and covariates, yet leave the truncation form unspecified. We eliminate the truncation distribution using a pairwise likelihood argument, and construct a composite likelihood for the parameters of interest only. Simulation studies showed a substantial efficiency gain of the proposed method, especially for the regression coefficients.

In the second project, the PLAC estimator is extended to incorporate extraneous time-dependent covariates to study the association between time to death and treatment among patients with end-stage renal disease. The transplantation registry violates of the independence between the underlying truncation and covariates. However, the pairwise likelihood can be modified to accommodate such types of dependence, so that the resulting estimator is still consistent, asymptotically normal and more efficient than the conditional approach estimator, as long as there is heterogeneity in the covariates before enrollment.

In the third project, we identify homogeneous subgroups within unbalanced longitudinal data. Most clustering methods require pre-specified number of clusters and suffer from locally optimal solutions. An extension of the clustering using fusion penalty to longitudinal data is proposed. Alternative formulation using mixed effect model with quadratic penalty on the random effects is considered to achieve more stable estimates. Simulations show the proposed method has robust performance under various magnitudes of within-cluster heterogeneity and random error. It performs better than the existing methods when the observations are sparse.

CHAPTER I

Introduction

Two types of outcomes naturally arise when a cohort are followed over a period of time. First, repeated measures on different characteristics of the subjects are collected. Second, the time taken until the event of interest, i.e., the survival time, is also recorded (Kalbfleisch and Prentice, 2002). These two types of outcomes usually interrelate with each other, since they reflect different aspects of the same unobserved underlying biological processes. When these outcomes are obtained from observation studies, analysis often faces greater challenges compared with those from well-designed experiments. Recognizing these challenges and offering robust and efficient statistical methods for the observational data constitute the main focus of this dissertation.

One defining characteristic of survival data is that the outcomes could be incompletely observed. Right censoring and left truncation are the most common incompleteness (Mandel, 2007). For instance, in the natural history of disease, the survival time is typically the duration from the disease onset to death. Ideally, an incident cohort of disease-free subjects should be recruited and followed till some subjects develop the disease and experience the failure event. Right censoring occurs when a patient is still event-free at the end of the follow-up, thus we only know

that the actual survival time is longer than the observed censoring time. When the disease is rare, however, in order to accumulate enough observed event times, a prevalent cohort consisting of diseased subjects who have not had the failure event at recruitment is preferred for cost efficiency and logistic consideration. In addition to right censoring, event times in a prevalent cohort are subject to delayed entry or left truncation. Unlike right-censored subjects, from which partial information about the survival can be obtained, left-truncated subjects have no chance to be sampled, thus their survival information cannot be revealed from the data. In this sense, left truncation is a special type of biased sampling; the population of interest also includes those who had the disease but died before the recruitment.

Longitudinal data are valuable for studying either the pathological course of a disease or the normative biological aging process. For responses varying with time, repeated measures taken on the same subjects contain richer information than the same amount of cross-sectional observations from different subjects. Nevertheless, longitudinal data on human subjects in epidemiology studies are almost always sparse, i.e., each subject only has a few follow-ups. Methods in functional data analysis, where the data is usually sampled over a fine time grid, are not directly applicable. Different subjects often have different observation times, that is, the data are irregular or unbalanced, which exclude most multivariate analysis tools in the analysis of such data. Even though the observations by design are separated by roughly regular intervals, using a different time scale (say, age of the subject) will make the data unbalanced. Moreover, longitudinal data are often measured with errors, which also adds to the difficulty of analysis.

In Chapter III, we consider efficiency improvement in regression methods for left-truncated data with additional distributional assumptions on the truncation times.

Conventional conditional approaches would correct the selection biases caused by truncation, yet may be inefficient due to ignoring the marginal information. Assuming parametric forms and modeling truncation times explicitly will bring considerable efficiency gain, yet the inferences could be misleading when the parametric forms are misspecified. To avoid restrictive parametric assumptions to still incorporate the additional marginal information, we proposed a pairwise likelihood augmented estimator for the Cox model (Cox, 1972). A pairwise pseudo-likelihood is used to eliminate the unspecified truncation distribution, and then combined with the conditional likelihood to form a composite likelihood for the parameters of interest. Simulation studies showed that the efficiency gain using the proposed method is substantial, especially under scenarios with shorter follow-up period and thus higher censoring rates. Appealing asymptotic properties of the proposed estimator including a closed-form consistent variance estimator are provided using empirical process and U-process theories.

Motivated by the United Network for Organ Sharing (UNOS) kidney transplantation registry data, in Chapter IV, the pairwise likelihood augmented Cox (PLAC) estimator is extended to cases where time-dependent covariates present. Although survival data involving both truncation and time-dependent covariates are ubiquitous in practice, careful investigation of the corresponding regression methods is rare in literature. Because estimating the effect of the time-dependent covariates requires fully-observed covariates history, the lack of information before enrollment for the prevalent cohort often hinder analysis which accounts for truncation. In stead, the issue is circumvented by selecting the enrollment time as the time of origin, which is not only less meaningful, but also incorrect in some cases (Sperrin and Buchan, 2013). The difficulty we faced in the UNOS data to apply the PLAC estimator is the

violation of the independence assumption between the covariates and the underlying truncation times. With a modification of the pairwise likelihood, we show that it can accommodate certain types of such dependence, including that in the UNOS data. The resulting modified estimator is still consistent, asymptotically normal and more efficient than the corresponding conditional approach estimator as long as there is heterogeneity in the time-dependent covariates before enrollment.

In Chapter V, we identify subgroups and structural patterns within sparse and irregular longitudinal trajectories. Common clustering methods usually require pre-specified number of clusters and suffer from locally optimal solutions. Convex clustering reformulates clustering as an optimization problem with fusion penalty on pairwise differences, which yields continuous clustering path and guarantees a unique global optimizer. An extension of the convex clustering to longitudinal data by solving a penalized least squares problem is provided. Quadratic penalty on the random effects to achieve more stable estimates is investigated. Simulations show the proposed method has good performance under various within-cluster heterogeneity and measurement errors, and it is more robust to the sparsity of the observations compared with the existing methods. Application to selected continuous outcomes from the NAS study is used to illustrate the usage of the proposed method.

The rest of the dissertation is organized as follows. Literature review for the related methods for all the projects is given in Chapter II. The body of this dissertation, consists of Chapter III through Chapter V introduces the proposed methodologies. Conclusions, discussions, and suggestions for future research are provided in Chapter VI. The appendices contain detailed asymptotic proofs, additional simulation results and data analysis results, followed by the bibliography.

CHAPTER II

Literature Review

In this chapter, we give some background for the methodologies covered in Chapter III through V. First, a survey of the length-biased sampling methods, a class of methods to improve estimation efficiency for left-truncated data under an additional uniform assumption on truncation times, is provided in Section 2.1. Although the distributional assumptions are different, the ideas behind these methods are similar to what our proposed method, the pairwise likelihood augmented estimator, relies on. Second, the theory of composite likelihood inferences are reviewed in Section 2.2, of which the pairwise pseudo-likelihood is a special case. Lastly, Section 2.3 gives an overview of existing methods and softwares for longitudinal data clustering; strengths and drawbacks of these clustering methods are highlighted.

2.1 Length-Biased Sampling Methods

The history of length-biased sampling can be traced back to Wicksell's corpuscle problem (Wicksell, 1925) in stereology. It was systematically studied in point processes (McFadden, 1962), electron tube life (Blumenthal, 1967), cancer screening trials (Zelen and Feinleib, 1969), and fiber length distribution (Cox, 1969). Under length-biased sampling, the probability of a unit being sampled is proportional to its length, size or other positive measures. In a prevalent cohort, if we assume the

disease incidence follows a stationary Poisson process (which usually holds for stable diseases), then the probability of a patient being sampled is proportional to his or her survival time (Shen et al., 2016). In this sense, length-biased sampling is a special form of left truncation under the stationarity assumption. Since the stationarity assumption implies that truncation times are uniform distributed, it is also referred to as the uniform truncation assumption.

Denote the independent underlying survival time and truncation time as T^* and A^*. In a prevalent cohort, only subjects with $(T, A) = (T^*, A^*) \mid (T^* > A^*)$ can be observed. The residual survival time after recruitment, denoted by V, is subject to right censoring by C, which is independent of (A, T). Let $X = \min(T, A + C)$, $\Delta = I(T \leqslant A + C)$. We use f, F and S to denote the density, distribution and survival functions of T^*, and the distribution function of A^* is denoted as G with density g. Under length-biased sampling, g is a constant, thus the joint density of (A, T) is $f(t)I(0 < a < t)/\mu$, where $\mu = \int_0^\infty S(a)da$ is the mean survival time. Denote by $\tilde{F}$ the distribution of the biased survival time T, then its density is given by $\tilde{f}(t) = tf(t)/\mu$ (Cox, 1969). In the renewal theory, A and V are referred to as backward and forward recurrence time, respectively. Under length-biased sampling, A and V share the same marginal density function. To see this, note that the joint distribution of (A, V) is given by $f(a + v)I(a > 0, v > 0)/\mu$. By integration,

$$f_V(t) = f_A(t) = \frac{S(t)}{\mu} I(t > 0). \tag{2.1}$$

When the truncation distribution is unspecified, the truncation product-limit estimator by conditioning on the truncation times is fully efficient (Wang, 1991). However, for length-biased sampled data, the product-limit estimator is inefficient, since it does not appreciate the known truncation distribution. The non-parametric maximum likelihood estimate (NPMLE) of F under length-biased sampling was first given

by Vardi (1982) when right-censoring is not allowed. Later, Vardi (1989) developed an expectation-maximization (EM) algorithm to estimate $\tilde{F}$ when right censoring presents, and F is obtained using back-transformation. In Vardi (1989), the so-called 'multiplicative censoring' is a specific form of informative censoring induced by the length-biased sampling scheme. It is worth noting that Vardi's NPMLE is characterized by jumping at both uncensored and censored times. Not like the product-limit estimator (Wang, 1991), Vardi's NPMLE does not have a closed-form expression. Huang and Qin (2011) proposed a non-parametric estimator of F, retaining the form of the product-limit estimator at the cost of a small efficiency loss compared with the NPMLE. Specifically, by (2.1), they calculate the Kaplan-Meier estimator $\tilde{S}_A$ with the pooled data $(A_i, \Delta_i = 1)$ and (V_i, Δ_i), $i = 1, \ldots, n$, and then plug it in the original product-limit estimator. Their estimator is shown to be more efficient than the product-limit estimator with a closed-form covariance matrix.

Let $\mathbf{Z}$ be a $p \times 1$ vector of covariates, and $\boldsymbol{\beta}$ the corresponding regression coefficients. Under length-biased sampling, individuals in the risk set $Y(x_i) \equiv \{j : x_j \geqslant x_i\}$ would have unequal probabilities to fail at x_i, even after adjustment by $\exp(\boldsymbol{\beta}^T \mathbf{Z}_j(x_i))$. Moreover, the standard partial likelihood approach under the Cox's model is inappropriate, since the full likelihood does not decompose the usual way. Wang (1996) proposed to construct unbiased risk sets at each x_i by sampling from $Y(x_i)$ and assigning less inclusion probability to larger x_j. Then one can construct a pseudo-likelihood similar to the Cox's partial likelihood:

$$\mathcal{L}^*(\boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\exp\left\{\boldsymbol{\beta}^T \mathbf{Z}_i(x_i)\right\}}{\sum_{j \in Y^*(x_i)} \exp\left\{\boldsymbol{\beta}^T \mathbf{Z}_j(x_i)\right\}}.$$

Wang (1996) also suggested replicating the procedure to remedy the extra variation introduced by the sampling. However, the method does not allow for right censoring, thus its practical use is limited.

Unbiased estimating equations are appealing alternatives when maximizing the full likelihood is difficult. Let H be an arbitrary increasing function and ε_T a random variable with known density. Shen et al. (2009) develop an unbiased estimating equation for length-biased data under the semi-parametric transformation model, $H(T^*) = -\boldsymbol{\beta}^T\mathbf{Z} + \varepsilon_T$:

$$U_T(\boldsymbol{\beta}) = \sum_{i,j} q(\mathbf{Z}_{ij})\mathbf{Z}_{ij}\delta_i\delta_j \left\{ \frac{I(X_i \geqslant X_j) - \xi(\boldsymbol{\beta}^T\mathbf{Z}_{ij})}{w_c(X_i)w_c(X_j)} \right\} = 0,$$

where $\mathbf{Z}_{ij} = \mathbf{Z}_i - \mathbf{Z}_j$, $w_c(y) = \int_0^y S_C(t)dt$, $\xi(\boldsymbol{\beta}^T\mathbf{Z}_{ij}) = \mathrm{E}(I(T_i^* \geqslant T_j^*)|\mathbf{Z}_i, \mathbf{Z}_j)$, S_C is the survival function of the censoring time, and $q(\cdot)$ is a positive weight function. Shen et al. (2009) also proposed estimating equations for the semi-parametric accelerated failure time model, $\log T^* = \boldsymbol{\beta}^T\mathbf{Z} + \epsilon_A$, where ϵ_A has an unknown distribution with mean zero:

$$U_A(\boldsymbol{\beta}) = \sum_{i=1}^{n} q(\mathbf{Z}_i)\delta_i\mathbf{Z}_i \frac{(\log X_i - \mathbf{Z}_i^T\boldsymbol{\beta})}{w_c(X_i)} = 0.$$

When S_C is unknown, the Kaplan-Meier estimator can be plugged in to get asymptotically unbiased estimating equations.

Under the Cox model, inverse weighted estimating equations were proposed by Qin and Shen (2010)

$$U_{C1}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{Z}_i - \frac{\sum_{j=1}^n I(X_j \geqslant X_i)\delta_j\mathbf{Z}_j \exp(\boldsymbol{\beta}^T\mathbf{Z}_j)\{X_j S_C(X_j - A_j)\}^{-1}}{\sum_{j=1}^n I(X_j \geqslant X_i)\delta_j \exp(\boldsymbol{\beta}^T\mathbf{Z}_j)\{X_j S_C(X_j - A_j)\}^{-1}} \right] = 0\,;$$

$$U_{C2}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{Z}_i - \frac{\sum_{j=1}^n I(X_j \geqslant X_i)\delta_j\mathbf{Z}_j \exp(\boldsymbol{\beta}^T\mathbf{Z}_j)\{w_c(X_j)\}^{-1}}{\sum_{j=1}^n I(X_j \geqslant X_i)\delta_j \exp(\boldsymbol{\beta}^T\mathbf{Z}_j)\{w_c(X_j)\}^{-1}} \right] = 0,$$

where S_C and w_c are defined as above. The estimates of this estimating equation can be obtained by fitting common Cox model with appropriate weights.

Nevertheless, estimating equations are and in general less efficient than the corresponding maximum likelihood approaches. Moreover, all the above estimating

equations require S_C, and hence might be less robust against different censoring distributions. Qin et al. (2011) proposed an expectation-maximization (EM) algorithms to jointly estimate the $\lambda_0(t)$ and $\boldsymbol{\beta}$ from the full likelihood of length-biased data under the Cox model. Unlike Vardi (1989), the 'missing data' in their EM algorithm treats are the latent truncated subjects, and the algorithm directly estimates the unbiased distribution F in stead of $\tilde{F}$.

Although the modified $\Lambda_0(t)$ has a closed form in Qin et al. (2011), the EM algorithm is computation-intensive. Note that the full likelihood can be decomposed:

$$\begin{aligned}\mathcal{L}_n(\beta,\Lambda) &\propto \prod_{i=1}^{n} \frac{f(X_i|\mathbf{Z}_i)^{\delta_i} S(X_i|\mathbf{Z}_i)^{1-\delta_i}}{\mu(\mathbf{Z}_i)} = \prod_{i=1}^{n} \frac{f(X_i|\mathbf{Z}_i)^{\delta_i} S(X_i|\mathbf{Z}_i)^{1-\delta_i}}{S(A_i|\mathbf{Z}_i)} \times \prod_{i=1}^{n} \frac{S(A_i|\mathbf{Z}_i)}{\mu(\mathbf{Z}_i)} \\ &= \mathcal{L}_n^C(\beta,\Lambda) \times \mathcal{L}_n^M(\beta,\Lambda).\end{aligned}$$

To avoid the high-dimensional optimization, a maximum pseudo-profile likelihood estimator (MPPLE) was proposed by Huang et al. (2012). The Breslow estimator from $\mathcal{L}_n^C$ is plugged into $\mathcal{L}_n$ to obtain a pseudo likelihood for $\boldsymbol{\beta}$ only.

Huang and Qin (2012) proposed a composite partial likelihood (CPL) method to for length-biased data under the Cox model. The proposed method relies on (2.1), and is closely related to the estimator in Huang and Qin (2011). Assuming $\Delta_i = 1$ for $i = 1, \ldots, m$ and $\Delta_i = 0$ for $i = m+1, \ldots, n$. A composite likelihood is constructed as the product of the conditional likelihood of V given A and that of A given V.

$$\mathcal{L}_n^C = \prod_{i=1}^{m} \left\{ \frac{f(X_i|\mathbf{Z}_i)}{S(A_i|\mathbf{Z}_i)} \times \frac{f(X_i|\mathbf{Z}_i)}{S(V_i|\mathbf{Z}_i)} \right\} \times \prod_{i=m+1}^{n} \frac{S(X_i|\mathbf{Z}_i)}{S(A_i|\mathbf{Z}_i)}.$$

Profiling out Λ results in the composite partial likelihood

$$\mathcal{L}_n^{CP} = \prod_{i=1}^{n} \left\{ \frac{2\exp(\beta^T\mathbf{Z}_i)}{\sum_{j=1}^{n} \exp(\beta^T\mathbf{Z}_j)\left(I(A_j \leqslant X_i \leqslant X_j) + \Delta_j I(V_j \leqslant X_i \leqslant X_j)\right)} \right\}^{2\Delta_i}.$$

The maximizer of $\mathcal{L}_n^{CP}$ is equivalent to the maximum partial likelihood estimator using the augmented data pooling $\{(X_i, A_i, \mathbf{Z}_i, \Delta_i)\}_{i=1}^n$ and $\{(X_i, V_i, \mathbf{Z}_i, \Delta_i = 1)\}_{i=1}^m$.

All length-biased sampling methods rely on the stationarity assumption, which is crucial to check as a model diagnostic step. Asgharian et al. (2006) provided a simple graphical checking method. By (2.1), we can plot the K-M estimators for both A and V, and check for discrepancy. Mandel and Betensky (2007) provided formal tests for the uniform truncation. One of the goodness-of-fit tests is closely related to the multiplicative censoring (Vardi, 1989). Let $Q = A/T$ with distribution F_Q, then $F_Q = U(0,1)$ if and only if G is uniform. They therefore suggested to compare $\hat{F}_Q$ to $U(0,1)$ using a one-sample Kolmogorov-Smirnov test. By applying the inverse probability transformation, we can actually test $H_0 : G = G_0$, for any known continuous G_0. However, this test can only be used on the uncensored data, since the test statistic depends on T_i's. When there is censoring, weighted log-rank tests for paired censored data Jung (1999) for the equality the distributions of A and V can be used, which formalizes the graphical method by Asgharian et al. (2006).

Papers on methods for length-biased sampled data keep emerging in the literature in recent years (Asgharian et al., 2014; Shen et al., 2016). Similar to the case-control sampling, length-biased sampling is a form of outcome-dependent sampling. The stationarity assumption holds at least approximately in various applications (Asgharian et al., 2002; de Uña-álvarez, 2004). Actually, even when G is parametrically modeled or even left completely unspecified, the idea of retrieving information from the marginal likelihood to improve efficiency can still be adopted (Liu et al., 2016; Huang and Qin, 2013). An specific approach under the independence assumption between the underlying truncation times and the covariates, the pairwise likelihood augmented estimator, will be introduced in Chapter III and IV.

2.2 Composite Likelihood Methods

The full likelihood approach as well as the corresponding maximum likelihood estimator (MLE) is often considered as the gold standard in statistical inferences. The MLE has the merits such as consistency, asymptotic normality and asymptotic efficiency. However, correct specification of the full likelihood is not always an easy task. Even when the specification is straightforward, the tremendous computation burden of maximizing a cumbersome full likelihood will often make the model infeasible in practice. To this end, alternatives based on modification of the full likelihood have been proposed during the past four decades.

Let Y be an $m \times 1$ random vector with joint density $f(y;\theta)$, where the parameter $\theta \in \Theta \subseteq \mathbb{R}^p$. Denote by $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ a set of marginal or conditional events with associated likelihoods $\mathcal{L}_k(\theta; y) \propto f(y \in \mathcal{A}_k; \theta)$. A composite likelihood of the K events is defined as the weighted product

$$\mathcal{L}^c(\theta; y) = \prod_{k=1}^{K} \mathcal{L}_k(\theta; y)^{\omega_k}, \tag{2.2}$$

where ω_k are non-negative weights. The idea of composite likelihood dates back to Besag (1974) in spacial statistics, while the term was coined by Lindsay (1988) which describes the nature of this class of pseudo-likelihoods. Comprehensive overviews on this topic can be found in Varin (2008) and Varin et al. (2011).

Based on the form of the likelihood objects in (2.2), composite likelihoods are divided into composite conditional likelihoods and composite marginal likelihoods. The pseudo-likelihood of Besag (1974) and the partial likelihood (Cox, 1975) are examples of the composite conditional likelihoods. They share the characteristics of omitting terms which complicate the evaluation of the full likelihood yet contain little information for the parameters of interest. On the other hand, when the focus is

on the marginal mean and/or dependence structure, composite marginal likelihoods usually consist production of low-dimensional marginal densities. Examples include pseudo-likelihoods constructed under working independence, pairwise likelihoods or combination of the two (Cox and Reid, 2004).

The maximum composite likelihood estimator (MCLE) $\hat{\theta}^c$ maximize (2.2), or its logarithm $\ell^c(\theta; y) = \sum_{k=1}^{K} \ell_k(\theta; y) w_k$, where $\ell_k(\theta; y) = \log \mathcal{L}_k(\theta; y)$. The MCLE may be found by solving the composite score function $U(\theta; y) = \nabla_\theta \ell^c(\theta; y)$, which is a linear combination of the scores associated with each component log-likelihood $\ell_k(\theta; y)$. Let $Y_1, \ldots, Y_n$ be independently and identically distributed (i.i.d.) random vectors from $f(y, \theta)$. Under regularity conditions, because $U(\theta; y)$ is the linear combination of the score functions corresponding to $\ell_k(\theta; y)$, $\hat{\theta}^c$ is consistent. Furthermore, additional smoothness conditions of the composite likelihood score statistic and the central limit theorem lead to $\sqrt{n}(\hat{\theta}^c - \theta) \rightsquigarrow N_p(0, \mathcal{I}^{-1}(\theta))$, where $\mathcal{I}(\theta) = J(\theta)^{-1} V(\theta) J(\theta)^{-1}$ is the Godambe information matrix for a single observation. The sensitivity matrix $J(\theta) = \mathrm{E}_\theta\{-\nabla_\theta u(\theta, Y)\}$, whereas the variability matrix $V(\theta) = \mathrm{Var}_\theta\{u(\theta, Y)\}$. Analogous to MLE under model misspecification, efficiency loss comparing to the full likelihood approach is expected (Kent, 1982; Lindsay, 1988; Molenberghs and Verbeke, 2005).

When a $q \times 1$ sub-vector ψ of the parameter θ is of interest, Wald and score test statistics following the usual asymptotic χ^2_q distribution for $H_0 : \psi = \psi_0$ can be constructed similarly from the composite likelihood (Molenberghs and Verbeke, 2005). Although likelihood ratio test might be preferable for its invariance under reparametrization and numerical stability, the test statistic from a composite likelihood has a non-standard asymptotic distribution involving a linear combination of χ^2_1 distributions (Kent, 1982). Estimation of $\mathcal{I}(\theta)$, especially $V(\theta)$ is computationally

demanding in constructing the test statistics. When the sample size is small compared with the dimension of the parameter, resampling methods such as jackknife (Zhao and Joe, 2005) or bootstrap can be used.

Model selection under composite likelihood can be conducted using information criteria. Varin and Vidoni (2005) proposed a generalized Takeuchi's information criterion. Unless the composite likelihood takes the form of an ordinary likelihood, it which will not reduce to Akaike's information criterion (AIC) even when the information equality holds. Gao and Song (2010) developed a composite likelihood version of Bayesian information criterion (CL-BIC) to get more parsimonious models. Xue et al. (2012) extended penalized likelihood estimation to the sparse Ising models for complex interactions in network, where they used a composite conditional likelihood with penalty to get rid of the computationally intractable partition function. Using composite conditional likelihoods to eliminate quantities in the likelihood that is hard to estimate turns to be very useful, as will be shown later in the derivation of the pairwise likelihood in Chapter III.

The EM algorithms based on the composite likelihood have been investigated, of which Liang and Yu (2003) in network tomography is one earliest example. Recently, Gao and Song (2011) give a general composite marginal likelihood EM algorithm. Although composite likelihood methods are usually taken as a possible competitor within the frequentist framework for Markov Chain Monte Carlo (MCMC) methods, Bayesian inferences from composite likelihood (Ribatet et al., 2009) have also been proposed.

Both the composite likelihood and the generalized estimating equations (GEE) emerge when the full-likelihood inferences are intractable. To remedy the same issue, GEE replaces the score equations with estimating equations specifying of the

first two moments only, whereas the composite likelihood methods directly replaces the full likelihood with pseudo-likelihoods constructed by simpler components. Both approaches yield consistent and asymptotically normal estimators. In term of efficiency, Aerts et al. (2002) found that the composite marginal likelihood (CML) is very similar to GEE2, while its computational complexity is closer to GEE1. Moreover, Constructing test statistics invariant to parametrization and model-selection criteria are pretty straightforward for composite likelihood, but are difficult with estimating equations approaches. In handling complex and high-dimensional models, Varin (2008) suggested that the product form of CML eases the use of parallel implementation, and the corresponding results are easier to reproduce compared with simulation-based MCMC methods.

Composite likelihood methods find most of their usage in clustered and longitudinal data, time series, spacial data, genetics and multivariate survival analysis, where complicated dependent structures often arise. The use of lower-dimensional margins helps avoid high-dimensional matrix inversions and integrals (Renard et al., 2004; Bellio and Varin, 2005), thus the computation burden is substantially reduced. Specifically, the composite likelihood methods are useful in modeling mixture of continuous and discrete outcomes (De Leon, 2005; De Leon et al., 2007). Molenberghs and Verbeke (2005) devotes several chapters on the composite likelihood methods in longitudinal data (see, e.g., Chapter 9, 12, 21 and 24). Parzen et al. (2007) proposed a pairwise likelihood approach for longitudinal binary data with non-monotone non-ignorable missingness. Bruckers et al. (2016) used a pairwise likelihood to solve the model-based clustering of multivariate longitudinal data.

In spite of the growing interest and the appealing features of the composite likelihood methods, they are not panacea. A list of open questions exist in the area,

including but not limited to, how to make choices when different composite likelihoods are possible; how to select optimal weights if we are to combine several composite likelihoods; and which subset of parameters is identifiable. The extent to which the efficiency is lost comparing to full likelihood approach loses is also not well quantified. In large p small n problems, the consistency of composite likelihood methods is not clear. Besides all these concerns of the theories, there is also a lack of standard softwares for general use to implement composite likelihood methods.

2.3 Clustering Methods for Longitudinal Data

Cluster analysis identifies groups (clusters) in the sample without knowing the group labels a priori. A tremendous amount of literature exists on clustering in pattern recognition, machine learning and statistics (Jain, 2010). In this review, we will emphasize clustering methods suitable for longitudinal data, either heuristic ones or more formal ones based on models.

Longitudinal data are sometimes referred as sparse functional data. Jacques and Preda (2014) categorized the existing methods for functional data clustering into four groups: raw-data methods, filtering methods, adaptive methods and distance-based methods. James and Sugar (2003) is among the earliest for model-based clustering methods which takes functional data point of view on longitudinal data. They considered two types of likelihoods:

$$L_C(\theta_1, \ldots, \theta_G; z_1, \ldots, z_G | x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{z_i}(x_i | \theta_{z_i}) \tag{2.3}$$

and

$$L_M(\theta_1, \ldots, \theta_G; \pi_1, \ldots, \pi_G | x_1, \ldots, x_n) = \prod_{i=1}^{n} \sum_{k=1}^{G} \pi_k f_k(x_i | \theta_k). \tag{2.4}$$

For each subject, the (2.3) allocates a unique cluster membership, whereas (2.4) assigns a multinomial distribution on all possible clusters. By projecting onto the

natural cubic spline basis, they proposed the following functional clustering model:

$$\mathbf{Y}_i = \mathbf{S}_i(\lambda_0 + \Lambda\alpha_{z_i} + \gamma_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon_i \sim N(0, \sigma^2 I), \quad \gamma_i \sim N(0, \Gamma), \tag{2.5}$$

where $\mathbf{S}_i = (\mathbf{s}(t_{i1}), \ldots, \mathbf{s}(t_{in_i}))^T$ is the basis matrix, subject to the constrains $\sum_k \alpha_k = 0$ and $\Lambda^T \mathbf{S}^T \Sigma^{-1} \mathbf{S} \Lambda = I$. Fitting (2.5) under either (2.3) or (2.4) involves an iterative EM procedure. Beside the algorithm, James and Sugar (2003) also provide a series of clustering tools including low-dimensional representation of the curves, regions of greatest separation and curve prediction. Their approach can also be generalized to incorporate multiple curves and time-invariant covariates, with a similar formulation.

Chiou and Li (2007) viewed each curve is sampled from a mixture of stochastic processes in $L^2(\mathcal{T})$ associated with a random cluster C:

$$Y^{(c)}(t) = \mu^{(c)} + \sum_{j=1}^{\infty} \xi_j^{(c)} \rho_j^{(c)}, \qquad \text{for} \quad C = c. \tag{2.6}$$

If Y belongs to c, then (2.6) is its Karhunen-Leéve expansion, otherwise, discrepancies should exist. In this case, only the relative likelihood of the clusters matters, which gives rise to an induced distance measure. At each iteration, given the current clusters, the criterion $c^*(y) = \text{argmin}_c \|y - \tilde{y}^{(c)}\|$ is used to reclassify the curves, which results in a functional version the well-known k-means algorithm. Simulation (on a regular design) shows this method is comparable or superior to (2.5) by taking into account the modes of variation in addition to the mean structure. However, their approach will performance will be quite poor, unless the observations are densely observed on a regular grid.

The most important ingredient in clustering analysis is a suitable distance. Peng and Müller (2008) defined a distance for sparse and irregular data based on the idea of conditional expectation (Yao et al., 2005). Let $D(i, j)$ be the common L_2 distance between two curves $g_i(t)$ and $g_j(t)$. Assuming Y_i and Y_j are their observations, then

for the sparse and irregular observations, define the distance between them as

$$\tilde{D}(i,j) = \left\{E(D^2(i,j)|Y_i, Y_j)\right\}^{1/2} = \left\{E\left(\sum_{k=1}^{\infty}(\xi_{ik} - \xi_{jk})^2|Y_i, Y_j\right)\right\}^{1/2}. \tag{2.7}$$

Furthermore, define the truncated version $\tilde{D}^{(M)}(i,j)$ by Karhunen-Leéve expansion. By construction, these conditional expectations are random variables depending on the observed data, and are unbiased estimator for the corresponding distances. With this distance, Peng and Müller (2008) then conducted multidimensional scaling to project the curves on to a Euclidean spaces with lower dimension, and then used common k-means on the projection. The estimation of the $\tilde{D}^{(M)}(i,j)$ is done by solving penalized least-square problems for ξ_i's

Actually, given the principal components for the longitudinal data by Yao et al. (2005), various multivariate data clustering methods are available to cluster the principal component coefficients. Because the principal components are selected in the way that reserves the most variability (information) of the original data, we expect the vectors used in this filtrating approach (following the terminology by Jacques and Preda (2014)) would give a reasonably good surrogate of the raw observations. For example, a popular model-based approach would be Gaussian mixture model Biernacki et al. (2006).

Besides the above methods with a functional data point of view, there are other authors who take more conventional approaches such as mixture mixed effect models. In psychology studies, the clustering of longitudinal trajectories are often accompanied with post clustering analysis involving relate the cluster membership to the other characteristics of the subjects in the same group (Nagin and Odgers, 2010). These approaches are called 'group-based methods by psychologists, which are the methods behind the SAS procedure PROC TRAJ (SAS Institute Inc., 2011). An

effort to extend the heuristic k-means method is proposed in Genolini and Falissard (2010) and latter extended to cases where multiple trajectories are to be clustered (Genolini et al., 2013). However, their k-means methods can only be applied to balanced longitudinal data, with some imputation methods suggested to account for possible missing data in the repeated measures. Lastly, another model-based longitudinal data clustering methods was proposed by McNicholas and Murphy (2010), where a modified Cholesky decomposition was employed to get parsimonious model of the covariance structure.

Most existing longitudinal data clustering methods, like clustering methods for multivariate data, require pre-specified number of clusters and usually will have the problem of converging to local optimal solutions. Convex clustering, which can be seen as a convex relaxation of the k-means clustering or hierarchical clustering, is proposed by several authors aiming to solve these issues (Lindsten et al., 2011; Hocking et al., 2011). The method we proposed in Chapter V can be seen as an extension of the convex clustering to sparse and irregular longitudinal observations.

CHAPTER III

A Pairwise Likelihood Augmented Cox Estimator for Left-Truncated Data

3.1 Introduction

Survival data collected from a prevalent cohort, who already have the disease under study at enrollment, are subject to left-truncation. This is because those who died with the disease before enrollment would have no chance to be selected, whereas the selected patients, having survived until the enrollment, are healthier on average. To avoid overestimating the survival, conventional approaches make inferences conditional on truncation times (Kalbfleisch and Lawless, 1991; Wang et al., 1993). These approaches disregard the information about the regression coefficients in the marginal likelihood of the truncation times, and hence loss of efficiency is expected when additional knowledge on the underlying truncation distribution is available (Huang and Qin, 2012).

If the underlying truncation time is uniform distributed, left-truncation reduces to length-biased sampling (Vardi, 1989), that is, the selection probability of a subject is proportional to the length of his or her underlying survival time. A recent review paper by Shen et al. (2016) summarizes the non- and semi-parametric methods in the existing literature for length-biased data. Among the newly developed regression methods for length-biased data, many show considerable improvement of efficiency in

estimation compared with the conditional approach by incorporating the information in the marginal likelihood of the truncation times (Qin and Shen, 2010; Qin et al., 2011; Huang and Qin, 2012; Huang et al., 2012; Ning et al., 2014). Nevertheless, when the uniform truncation assumption is violated, these methods may yield inconsistent estimates (Huang and Qin, 2012).

The motivating study is a prevalent cohort study of patients with chronic kidney disease (CKD), sponsored by the Renal Research Institute (Perlman et al., 2003). Following the diagnosis, in general, CKD patients are referred to nephrologists to receive special care and treatments. The investigators were interested in whether the patient characteristics at referral were related to the disease progression to end-stage renal disease (ESRD) or death. At the study recruitment from June 2000 to January 2006, subjects with glomerular filtration rate (GFR) less than or equal to 50 ml/min/1.73 m^2 were invited to participate. The dataset is of a moderate sample size, so improving the estimation efficiency is important. However, statistical assessment in Section 3.4 indicated deviation of the motivating data from the uniform truncation assumption, which prompted us to seek an efficiency-improving method avoiding the potential biases when using the methods proposed for length-biased data.

Recently, Huang and Qin (2013) proposed a more efficient estimator for the additive hazards model under general left-truncation. They used a pairwise likelihood of the truncation times to eliminate the unspecified truncation distribution (Liang and Qin, 2000). In practice, however, the Cox model is more commonly used than additive hazards model, and its interpretation is familiar to practitioners (Cox, 1972). Yet the challenge of applying the pairwise likelihood approach to the Cox model lies in the complicated way that the pairwise likelihood still involves the cumulative baseline hazard function, causing serious theoretical and computational difficulties.

In this chapter, we propose to augment the conditional likelihood with a pairwise likelihood constructed from the marginal likelihood of the truncation times to improve the efficiency in estimation for the Cox model. We have achieved several important improvements. First, we design an nonparametric maximum likelihood estimating procedure to estimate the cumulative baseline hazard function along with the regression coefficients. Second, with the asymptotic results proven by empirical process and U-process theories, we provide a closed-form consistent sandwich variance estimator. Finally, we provide an iterative algorithm that explores the self-consistency of the nonparametric estimator and guarantees a computationally efficient implementation. An R package, `plac`, implementing the proposed method is available on CRAN (R Core Team, 2016). Our simulations show that efficiency of both the regression coefficients and the cumulative baseline hazard function, especially the former, can be improved using the proposed method. Moreover, even when the uniform truncation assumption holds, the proposed estimator of the regression coefficients has efficiency comparable to that of the full maximum likelihood estimator (MLE) proposed by Qin et al. (2011), and enjoys smaller biases. Thus, we believe the proposed estimator provides a promising alternative to improve the estimation efficiency for left-truncated survival data.

3.2 Proposed Method

3.2.1 Preliminaries

Suppose we choose the disease onset as the time origin. For a patient from the target population, let T^* denote the *underlying* survival time, i.e. time to event, and A^* denote the *underlying* truncation time, i.e. time to study enrollment. We use f and S to denote the density and survival functions of T^*, and the distribution function of A^* is denoted as G. Let Z^* be a $p \times 1$ vector of covariates. We assume A^*

and T^* are independent conditional on Z^*. A commonly used model that links the hazard function of T^* to the covariates Z^* is the Cox proportional hazards model (Cox, 1972):

$$\lambda(t \mid Z^*; \beta) = \lambda(t) \exp(\beta^{\mathrm{T}} Z^*),$$

where $\lambda(\cdot)$ is an unspecified baseline hazard function, and β is a $p \times 1$ vector of regression coefficients. The cumulative baseline hazard function is defined as $\Lambda(t) = \int_0^t \lambda(s)\, ds$. Data collected from a prevalent cohort only consists of patients with $A^* \leqslant T^*$. The same notations without asterisks, A, T, and Z, will be used throughout the chapter to denote the *observed* random variables conditional on $A^* \leqslant T^*$, i.e., $(A, T, Z) \equiv (A^*, T^*, Z^*) | (A^* \leqslant T^*)$.

Usually, the survival time is also subject to potential censoring by C starting from the enrollment. Thus, what we can observe are $X = \min(A + C, T)$ and $\Delta = I(T \leqslant A + C)$, where $I(\cdot)$ is the indicator function. Suppose we have independent and identically distributed observations $\{(A_i, X_i, \Delta_i, Z_i);\ i = 1, \ldots, n\}$ on n individuals sampled from a prevalent cohort. The full likelihood of the observed data is proportional to

$$\prod_{i=1}^{n} \mathrm{pr}(A_i^*, T_i^*, C_i \mid Z_i^*, A_i^* \leqslant T_i^*) \propto \prod_{i=1}^{n} \frac{f(X_i \mid Z_i)^{\Delta_i} S(X_i \mid Z_i)^{1-\Delta_i} dG(A_i)}{\int_0^\infty S(a \mid Z_i) dG(a)} \equiv \mathcal{L}_n,$$

We assume C is independent of (A, T) given Z, and that A^* does not depend on Z^*, i.e., the underlying patient recruitment process does not depend on the covariates. Note that the latter assumption does not imply independence between the *observed* A and Z, since the biased sampling scheme may induce correlations between them as well as between A and T given Z. The full likelihood can be further decomposed into two parts:

$$\mathcal{L}_n = \prod_{i=1}^{n} \frac{f(X_i \mid Z_i)^{\Delta_i} S(X_i \mid Z_i)^{1-\Delta_i}}{S(A_i \mid Z_i)} \times \prod_{i=1}^{n} \frac{S(A_i \mid Z_i) dG(A_i)}{\int_0^\infty S(a \mid Z_i) dG(a)} \equiv \mathcal{L}_n^C \times \mathcal{L}_n^M, \tag{3.1}$$

where $\mathcal{L}_n^C$ is the conditional likelihood of (X_i, Δ_i) given (A_i, Z_i), and $\mathcal{L}_n^M$ is the marginal likelihood of A_i given Z_i, for $i = 1, \ldots, n$.

3.2.2 Pairwise-Likelihood Augmented Cox (PLAC) Estimator

In the presence of truncation, inference based on $\mathcal{L}_n^C$ only, using the Cox's partial likelihood (Cox, 1975) with the at-risk indicator $Y_i(t) = I(A_i \leqslant t \leqslant X_i)$, has been proposed by Kalbfleisch and Lawless (1991) and Wang et al. (1993). The conditional approach yields consistent estimates, but it may be inefficient when additional assumption can be made on the truncation distribution, since it completely ignores the information about the parameters contained in $\mathcal{L}_n^M$. Taking advantage of the fully specified uniform truncation distribution, regression methods for length-biased data generally result in more efficient estimators. Among these methods, the expectation-maximization (EM) algorithm by Qin et al. (2011) yields asymptotically efficient estimator for the Cox model under the uniform truncation assumption. Recently, Liu et al. (2016) extended the EM algorithm by Qin et al. (2011) to general biased-sampling cases, where G is known up to some unspecified finite-dimensional parameters, and the estimation efficiency of the Cox model can be improved while jointly estimating these truncation distribution parameters.

Deviating from most existing efficiency improving methods in the literature for left-truncated data, our method does not impose any parametric assumptions on the underlying truncation distribution, nor on the baseline hazard function. Our approach to improving efficiency is to supplement $\mathcal{L}_n^C$ with major information in $\mathcal{L}_n^M$ that depends on β and λ only. Specifically, we first apply the pairwise likelihood method by Liang and Qin (2000) to $\mathcal{L}_n^M$ in order to eliminate the truncation distribution function, and then estimate β and λ based on a composite likelihood consisting of $\mathcal{L}_n^C$ and $\mathcal{L}_n^P$, where the pairwise likelihood $\mathcal{L}_n^P$ is derived as follows.

Suppose a sample $\{(A_i, Z_i), (A_j, Z_j); i < j\}$ is available. The pseudo-likelihood of the pair (i, j), conditional on (Z_i, Z_j) and the order statistic of (A_i, A_j), is

$$\frac{\frac{S(A_i|Z_i)dG(A_i)}{\int_0^\infty S(a|Z_i)dG(a)} \times \frac{S(A_j|Z_j)dG(A_j)}{\int_0^\infty S(a|Z_j)dG(a)}}{\frac{S(A_i|Z_i)dG(A_i)}{\int_0^\infty S(a|Z_i)dG(a)} \times \frac{S(A_j|Z_j)dG(A_j)}{\int_0^\infty S(a|Z_j)dG(a)} + \frac{S(A_i|Z_j)dG(A_i)}{\int_0^\infty S(a|Z_j)dG(a)} \times \frac{S(A_j|Z_i)dG(A_j)}{\int_0^\infty S(a|Z_i)dG(a)}} = \frac{1}{1 + R_{ij}(\beta, \Lambda)},$$

where

$$R_{ij}(\beta, \Lambda) = \frac{S(A_i \mid Z_j)S(A_j \mid Z_i)}{S(A_i \mid Z_i)S(A_j \mid Z_j)} = \exp\left[(e^{Z_i^{\mathrm{T}}\beta} - e^{Z_j^{\mathrm{T}}\beta})\{\Lambda(A_i) - \Lambda(A_j)\}\right]$$

denotes the generalized odds ratio under the Cox model. The pairwise likelihood $\mathcal{L}_n^P$ of all pairs is then given by

$$\mathcal{L}_n^P = \prod_{i<j} \{1 + R_{ij}(\beta, \Lambda)\}^{-1}.$$

It is worth noting that, by canceling out the terms involving G, $\mathcal{L}_n^P$ is a function of (β, Λ) only, whereas $\mathcal{L}_n^M$ is a function of (β, Λ, G). An alternative approach would be to directly maximize the full likelihood $\mathcal{L}_n^C \times \mathcal{L}_n^M$ over (β, Λ, G), which may be more efficient than the composite likelihood approach. However, when G is completely unspecified, maximizing over infinite dimensional parameters in addition to Λ will increase computational cost and can be unstable numerically in the real data; thus, we will not attempt to estimate G when it is not a parameter of interest. Simulation studies (Qin and Liang, 1999; Liang and Qin, 2000) show that the pairwise likelihood can retain the majority of the information in the likelihood from which it is derived, and that the efficiency loss may not be substantial, depending on the model as well as the values of the parameters. Therefore, to estimate β and λ, we propose using $\mathcal{L}_n^P$ as a reasonably good surrogate for $\mathcal{L}_n^M$ in the full likelihood approach. The analogous idea has been exploited in the additive hazards model by Huang and Qin (2013); however, the additive hazards model is less commonly used. Applying the

pairwise-likelihood augmentation method to the Cox model will greatly promote more practical use due to ease of interpretation to practitioners.

To account for the different magnitudes of $\log \mathcal{L}_n^C$ and $\log \mathcal{L}_n^P$ (there are n terms in $\log \mathcal{L}_n^C$ and $n(n-1)/2$ terms in $\log \mathcal{L}_n^P$), we maximize the following composite log-likelihood function:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i\left\{\log\lambda(X_i)+Z_i^{\mathrm{T}}\beta\right\}-\exp(Z_i^{\mathrm{T}}\beta)\int_0^{\infty}Y_i(t)\lambda(t)dt\right]$$
$$-\frac{2}{n(n-1)}\sum_{i<j}\log\{1+R_{ij}(\beta,\Lambda)\},$$

over the domain of (β, Λ). Using the nonparametric maximum likelihood estimation approach, we treat $\Lambda(\cdot)$ as a nondecreasing step function with jumps, denoted by $\Lambda\{\cdot\}$, only at the time points where events are observed and $\Lambda(0) = 0$ (see Murphy et al., 1997; Zeng and Lin, 2006, among others). Let $w_1 < \cdots < w_m$, $m \leqslant n$, be the ordered distinct observed event times, and $\lambda_1 = \Lambda\{w_1\}, \ldots\ \lambda_m = \Lambda\{w_m\}$ be the corresponding positive jumps of Λ at these times. We denote by $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^{\mathrm{T}}$ the vector of all positive jumps. For $k = 0, 1, 2$, we define the following functions which appear in $\log \mathcal{L}_n^P$ and its derivatives:

$$Q_{ij}^{(k)}(t;\beta)=\left(Z_i^{\otimes k}e^{Z_i^{\mathrm{T}}\beta}-Z_j^{\otimes k}e^{Z_j^{\mathrm{T}}\beta}\right)\left\{I(t\leqslant A_i)-I(t\leqslant A_j)\right\},$$

where $Z^{\otimes 0} = 1$, $Z^{\otimes 1} = Z$, and $Z^{\otimes 2} = ZZ^{\mathrm{T}}$. Below we may suppress the dependence on model parameters, using R_{ij} and $Q_{ij}^{(k)}(t)$ to denote $R_{ij}(\beta, \Lambda)$ and $Q_{ij}^{(k)}(t;\beta)$ when the meanings of the notations are clear from the context. Replacing $\lambda(t)$ with $\Lambda\{t\}$, we modify the composite log-likelihood as a function of β and $\boldsymbol{\lambda}$:

$$\begin{aligned}\ell_n^c(\beta,\boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i\left(\log\Lambda\{X_i\}+Z_i^{\mathrm{T}}\beta\right)-\exp(Z_i^{\mathrm{T}}\beta)\sum_{k=1}^{m}\lambda_k Y_i(w_k)\right\}\\ &\quad -\frac{2}{n(n-1)}\sum_{i<j}\log\{1+R_{ij}(\beta,\boldsymbol{\lambda})\},\end{aligned}\tag{3.2}$$

where $R_{ij}(\beta, \boldsymbol{\lambda}) = \exp\{\sum_{k=1}^{m} \lambda_k Q_{ij}^{(0)}(w_k)\}$. We refer to the resulting maximizer $(\hat{\beta},\ \hat{\boldsymbol{\lambda}})$ (or equivalently $(\hat{\beta},\ \hat{\Lambda})$) as the pairwise likelihood augmented Cox (PLAC) estimator, where Λ at a fixed time point $t \in [0, \tau]$ is estimated by $\hat{\Lambda}(t) = \sum_{k=1}^{m} \hat{\lambda}_k\, I(w_k \leqslant t)$. Specifically, differentiating (3.2) with respect to $(\beta, \boldsymbol{\lambda})$ yields the composite score functions (the dependence on n is suppressed):

$$
\begin{aligned}
U_\beta(\beta, \boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n} Z_i \left\{ \Delta_i - e^{Z_i^{\mathrm{T}}\beta} \sum_{k=1}^{m} \lambda_k Y_i(w_k) \right\} - \frac{1}{n(n-1)} \sum_{i \neq j} \frac{\sum_{k=1}^{m} \lambda_k Q_{ij}^{(1)}(w_k)}{1 + R_{ij}^{-1}}, \\
U_{\lambda_k}(\beta, \boldsymbol{\lambda}) &= \frac{1}{n}\sum_{i=1}^{n} I(X_i = w_k) \left\{ \Delta_i/\lambda_k - Y_i(w_k)\, e^{Z_i^{\mathrm{T}}\beta} \right\} - \frac{1}{n(n-1)} \sum_{i \neq j} \frac{Q_{ij}^{(0)}(w_k)}{1 + R_{ij}^{-1}}.
\end{aligned}
$$

Let $U_{\boldsymbol{\lambda}}^{\mathrm{T}} = (U_{\lambda_1}, \ldots, U_{\lambda_m})$, then the PLAC estimator $(\hat{\beta}, \hat{\boldsymbol{\lambda}})$ is the solution to

$$
U(\beta, \boldsymbol{\lambda}) = (U_\beta^{\mathrm{T}}, U_{\boldsymbol{\lambda}}^{\mathrm{T}})^{\mathrm{T}}(\beta, \boldsymbol{\lambda}) = 0, \tag{3.3}
$$

which can be obtained numerically using the following algorithm, for example.

Unlike the conditional approach, directly solving the nonlinear system (3.3) is a difficult problem due to the computational complexity brought by the pairwise structure. Therefore, we propose the following algorithm to solve for $\hat{\beta}$ and $\hat{\lambda}_k$, $k = 1, \ldots, m$, iteratively:

Step 1. Start with initial values $\beta^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$.

Step 2. At the rth iteration, update each $\lambda_k^{(r)}$ using

$$
\lambda_k^{(r)} = \frac{\frac{1}{n}\sum_{i=1}^{n} \Delta_i I(X_i = w_k)}{\frac{1}{n}\sum_{i=1}^{n} Y_i(w_k) e^{Z_i^{\mathrm{T}}\beta^{(r-1)}} + \frac{1}{n(n-1)} \sum_{i \neq j} \frac{Q_{ij}^{(0)}(w_k; \beta^{(r-1)})}{1 + 1/R_{ij}(\beta^{(r-1)}, \boldsymbol{\lambda}^{(r-1)})}}. \tag{3.4}
$$

Step 3. Update $\beta^{(r)}$ by one step of Newton-Raphson iteration:

$$
\beta^{(r)} = \beta^{(r-1)} - \left\{ \dot{U}_{\beta\beta}(\beta^{(r-1)}, \boldsymbol{\lambda}^{(r)}) \right\}^{-1} \left\{ U_\beta(\beta^{(r-1)}, \boldsymbol{\lambda}^{(r)}) \right\},
$$

where $\dot{U}_{\beta\beta}(\beta^{(r-1)}, \boldsymbol{\lambda}^{(r)}) = \partial U_\beta(\beta, \boldsymbol{\lambda})/\partial \beta^{\mathrm{T}}|_{\beta=\beta^{(r-1)}, \boldsymbol{\lambda}=\boldsymbol{\lambda}^{(r)}}$.

Step 4. Repeat Steps 2 and 3 until convergence.

The initial values for the parameters in Step 1 can be set using $\beta^{(0)} = 0$ and $\boldsymbol{\lambda}^{(0)} = (1/m, \ldots, 1/m)$ or the estimates from the conditional approach. In our simulation studies, it is demonstrated that the algorithm is robust to the choice of initial values. In Step 2, updating λ_k using the self-consistent solution (3.4) is the crucial step which makes the computation of the PLAC estimator tractable in a reasonable amount of time. The above algorithm is implemented in the R package `plac` which is available on CRAN (R Core Team, 2016).

3.2.3 Asymptotic Properties

We establish the consistency and asymptotic normality of the PLAC estimator $(\hat{\beta}, \hat{\Lambda})$, utilizing techniques from both empirical process (van der Vaart and Wellner, 1996) and U-process theories (De la Peña and Giné, 1999). The asymptotic properties for the infinite-dimensional parameter Λ is proved on the interval $[0, \tau]$, where τ is the upper bound for the observed survival time $X = \min(A+C, T)$ (Qin et al., 2011). Denote the normalized score functions corresponding to $\mathcal{L}_n^C$ and $\mathcal{L}_n^P$ as $U^C(\beta, \boldsymbol{\lambda}) = n^{-1} \sum_{i=1}^n U_i^C(\beta, \Lambda)$ and $U^P(\beta, \boldsymbol{\lambda}) = 2\{n(n-1)\}^{-1} \sum_{i<j} U_{ij}^P(\beta, \boldsymbol{\lambda})$, respectively, where

$$U_i^C(\beta, \boldsymbol{\lambda}) = \begin{pmatrix} \Delta_i Z_i - Z_i e^{Z_i^{\mathrm{T}}\beta} \sum_{k=1}^m \lambda_k Y_i(w_k) \\ I(X_i = w_1)\{\Delta_i/\lambda_1 - Y_i(w_1)\, e^{Z_i^{\mathrm{T}}\beta}\} \\ \vdots \\ I(X_i = w_m)\{\Delta_i/\lambda_m - Y_i(w_m)\, e^{Z_i^{\mathrm{T}}\beta}\} \end{pmatrix} \tag{3.5}$$

and

$$U_{ij}^{P}(\beta, \boldsymbol{\lambda}) \;=\; -1/(1+R_{ij}^{-1}) \begin{pmatrix} \sum_{k=1}^{m} \lambda_k Q_{ij}^{(1)}(w_k) \\ Q_{ij}^{(0)}(w_1) \\ \vdots \\ Q_{ij}^{(0)}(w_m) \end{pmatrix}. \tag{3.6}$$

Theorem III.1 (Consistency)**.** *Under Conditions (C1)-(C4),*

$$\hat{\beta} \to \beta_0 \quad \textit{and} \quad \left\| \hat{\Lambda} - \Lambda_0 \right\|_{L_\infty[0,\tau]} \to 0 \qquad \textit{almost surely as } n \to \infty,$$

where $\|\cdot\|_{L_\infty[0,\tau]}$ *is the supremum norm on* $[0, \tau]$*.*

Under the regularity conditions specified in the Appendix A, Theorem III.1 shows that the PLAC estimator is a consistent estimator of the true parameters (β_0, Λ_0). The consistency proof follows three major steps. First, we show the parameters of interest (β_0, Λ_0) are identifiable. By the nature of the pairwise construction, $U_{ij}^{P}(\beta, \Lambda)$ is permutation-symmetric in the observed data; thus, the pairwise score function $U^{P}(\beta, \Lambda)$ and its derivatives are U-processes of order two. Second, we construct upper bounds for bracketing numbers of the related function classes by combining the bracketing entropy results of uniformly bounded monotone functions with the preservation theorems for Lipschitz function classes (see van der Vaart and Wellner, 1996, Chapter 2.7). The law of large numbers of these classes then follows from Corollary 3.2.5 of De la Peña and Giné (1999). In addition, we can show $\mathrm{E}\{U^{P}(\beta_0, \Lambda_0)\} = 0$ by the fact that $U_{ij}^{P}(\beta, \Lambda)$ is the exact score function corresponding to the pairwise likelihood of the pair (i, j), conditional on (Z_i, Z_j) and the order statistic of (A_i, A_j). In the last step, the strong consistency of the PLAC estimator can be proven through the likelihood equation argument similar to that given by Murphy et al. (1997), along with the composite Kullback-Leibler divergence (Varin and Vidoni, 2005) and the identifiability of the parameters.

For the weak convergence, we first establish the uniform $\sqrt{n}$-convergence rate and the asymptotic normality of the log-generalized odds ratio using the Hájek projection of U-processes (van der Vaart, 2000). The asymptotic normality of the PLAC estimator can be proved using Theorem 3.3.1 of van der Vaart and Wellner (1996). Noting that $n^{1/2}\,U(\beta_0, \Lambda_0) = n^{1/2}\,U^C(\beta_0, \Lambda_0) + n^{1/2}\,U^P(\beta_0, \Lambda_0)$, the asymptotic normality of $n^{1/2}\,U(\beta_0, \Lambda_0)$ is obtained by the separate contributions of $n^{1/2}\,U^C(\beta_0, \Lambda_0)$ and $n^{1/2}\,U^P(\beta_0, \Lambda_0)$, which are asymptotically independent (van der Vaart and Wellner, 1996, Example 1.4.6). The asymptotic normality of $n^{1/2}\,U^C(\beta_0, \Lambda_0)$ follows from the martingale theory (Andersen and Gill, 1982; Wang et al., 1993), and our innovative contribution is to identify the limiting distribution of $n^{1/2}\,U^P(\beta_0, \Lambda_0)$. The normality of the function classes involved in $U^P(\beta_0,\, \Lambda_0)$ and its derivative is shown through the results on the Vapnik-Chervonenkis subgraph classes, the normality of the log-generalized odds ratio, and the preservation theorems for Lipschitz functions (van der Vaart and Wellner, 1996, Chapter 2.10). Finally, the Fréchet-differentiability of $\mathrm{E}\{U(\beta_0,\, \Lambda_0)\}$ and the invertibility of its derivative can be shown by (C5) and the Fredholm theory, following arguments similar to those in Zeng and Lin (2006). The weak convergence results are summarized in the following theorem. Further detailed proofs are provided in Appendix A.

Theorem III.2 (Asymptotic normality)**.** *Under Conditions (C1)-(C4),* $n^{1/2}\,(\hat{\beta} - \beta_0,\, \hat{\Lambda}(t) - \Lambda_0(t))$ *converges weakly to a mean-zero Gaussian process in* $\mathbb{R}^p \times \mathrm{BV}[0, \tau]$*, where* $\mathrm{BV}[0, \tau]$ *denotes the space of all functions with bounded total variations on* $[0, \tau]$.

One of the appealing features of our approach is that the covariance of the limiting process of the PLAC estimator can be consistently estimated by a closed-form

sandwich estimator. Let

$$\begin{aligned}
\hat{V}^C &= \frac{1}{n}\sum_{i=1}^{n} U_i^C(\hat{\beta},\hat{\boldsymbol{\lambda}})^{\otimes 2},\\
\hat{J}^C &= -\frac{1}{n}\sum_{i=1}^{n} \partial U_i^C(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\lambda^{\mathrm{T}})\mid_{\beta=\hat{\beta},\,\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}},\\
\hat{V}^P &= \frac{4}{n-1}\sum_{i=1}^{n}\left\{\frac{1}{n-1}\sum_{i\neq j} U_{ij}^P(\hat{\beta},\,\hat{\boldsymbol{\lambda}})\right\}^{\otimes 2},\\
\hat{J}^P &= -\frac{1}{n(n-1)}\sum_{i\neq j} \partial U_{ij}^P(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})\big|_{\beta=\hat{\beta},\,\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}},
\end{aligned}$$

where the exact expressions of $\partial U_i^C(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})$ and $\partial U_{ij}^P(\beta,\boldsymbol{\lambda})/\partial(\beta^{\mathrm{T}},\boldsymbol{\lambda}^{\mathrm{T}})$ are given in the Web Supplementary. To define the asymptotic covariance, consider a linear functional

$$n^{1/2}\left[b_1^{\mathrm{T}}(\hat{\beta}-\beta_0)+\int_0^{\tau} h(t)d\left\{\hat{\Lambda}(t)-\Lambda_0(t)\right\}\right], \tag{3.7}$$

where $b_1\in\mathbb{R}^p$, and $h(t)$ is an arbitrary function with bounded total variation on $[0,\tau]$. Let b_2 be the $m\times 1$ vector $(h(w_1),\ldots,h(w_m))^{\mathrm{T}}$, and $b=(b_1^{\mathrm{T}},b_2^{\mathrm{T}})^{\mathrm{T}}$. For example, we can set $b_1=e_k, b_2=0$ when $\hat{\beta}_k, k=1,\ldots,p$, is of interest, where e_k is a unit vector with 1 at the k-th element and 0 otherwise. Whereas when $\Lambda(t)$ for a fixed t is the parameter of primary interest as in our simulation, we can set $b_1=0$ and $b_2=(I(w_1\leqslant t),\ldots,I(w_m\leqslant t))^{\mathrm{T}}$. As in Zeng and Lin (2006), since the PLAC estimator for Λ converges at a parametric rate, we can treat β and $\boldsymbol{\lambda}$ in (3.2) as if they are finite-dimensional parameters. Then by the asymptotic properties of U-statistics (Sen, 1960) and the composite likelihood theory, the linear functional (3.7) converges in distribution to a mean-zero Gaussian random variable with the variance that can be consistently estimated by $b^{\mathrm{T}}\hat{\Sigma}b$, where

$$\hat{\Sigma}=(\hat{J}^C+\hat{J}^P)^{-1}(\hat{V}^C+\hat{V}^P)(\hat{J}^C+\hat{J}^P)^{-1} \tag{3.8}$$

is the observed inverse Godambe information (Varin et al., 2011). Naturally, the sandwich estimator (3.8) has the following partition:

$$\hat{\Sigma} = \begin{pmatrix} \hat{\Sigma}_{\beta\beta} & \hat{\Sigma}_{\beta\boldsymbol{\lambda}} \\ \hat{\Sigma}_{\boldsymbol{\lambda}\beta} & \hat{\Sigma}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \end{pmatrix},$$

where the sub-matrices are the estimated asymptotic covariance matrices of the corresponding parameter estimates. The greatest advantage of (3.8) is that we can directly use the delta method to get the asymptotic variances of quantities of interest other than β and $\boldsymbol{\lambda}$. For instance, the asymptotic variance of $\hat{\Lambda}(t)$ can be estimated by

$$\hat{\Sigma}_{\hat{\Lambda}(t)} = \sum_{k=1}^{m} \sum_{l=1}^{m} I(w_k \leqslant t, w_l \leqslant t) \hat{\sigma}_{kl}^{(\boldsymbol{\lambda}\boldsymbol{\lambda})},$$

where $\hat{\sigma}_{kl}^{(\boldsymbol{\lambda}\boldsymbol{\lambda})}$ is the variance (covariance) estimate corresponding to λ_k and λ_l in $\hat{\Sigma}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}$.

3.3 Simulation

We conducted extensive simulation studies to evaluate the finite-sample performance of the proposed PLAC estimator, and compared it with the conditional approach estimator (Kalbfleisch and Lawless, 1991; Mandel and Betensky, 2007) and the MLE for length-biased data (LBML) proposed by Qin et al. (2011). The underlying survival time T^* was generated from a Cox model with two independent covariates:

$$\lambda(t \mid Z_1, Z_2; \beta_1, \beta_2) = \lambda(t) \exp(\beta_1 Z_1 + \beta_2 Z_2), \tag{3.9}$$

where $Z_1 \sim$ Binomial(0.5), $Z_2 \sim$ Uniform$(-1, 1)$, and the true values $\beta_1 = \beta_2 = 1$. The baseline hazard function was $\lambda(t) = 2t$. For the underlying truncation time, we considered two cases: (1) length-biased data with and (2) non-length-biased data with $A^* \sim$ Exp(1), i.e., exponential distribution with rate one. Note that under

length-biased sampling, the truncation distribution is uniform, thus the *observed* survival time T has density function $tf(t|Z)/\mu(Z)$, where $\mu(Z) = \int uf(u|Z)du$. In Case 1, we first generated the *observed* survival times t_i, $i = 1, \ldots, n$, and then drew corresponding *observe* truncation times a_i from Uniform$(0, t_i)$ (Mandel and Betensky, 2007). In Case 2, the underlying survival times t_i^* were generated from (3.9), and the underlying truncation times were from Exp(1); yet only the pairs (a_i^*, t_i^*) satisfying $a_i^* \leqslant t_i^*$ were kept until the desired sample size was reached. The censoring times c_i, $i = 1, \ldots, n$, were generated independently from Uniform$(0, C_{\max})$, where $C_{\max}$ was chosen to designate censoring rates of approximately 50% and 80%. The event indicator for subject i was obtained by calculating $\delta_i = I(t_i \leqslant a_i + c_i)$. Sample sizes of 200, 400 and 800 were considered, and we generated 1000 datasets under each scenario.

For each dataset, we estimated β_1, β_2, and $\Lambda(t)$ at two fixed times $t = (\tau_{30}, \tau_{60})$, where τ_{30} and τ_{60} were the 30% and 60% percentiles of the observed survival times under each scenario. Summary statistics for datasets with sample sizes of 400 and 800, including the average of the estimates minus the true value, the empirical standard error of the estimates, the average of the standard error estimates, the 95% coverage probability, and the relative efficiency (the ratio of the mean squared errors) relative to the conditional estimator, are provided in Table 3.1.

The empirical biases of the PLAC estimates, like the conditional approach estimates, are close to zero under all scenarios. When data are length-biased (Case 1), the LBML estimates also have biases that are small yet larger than those of the other two estimators. The moderate biases in LBML estimates has been consistently observed by Liu et al. (2016). In contrast, the maximum likelihood estimates in Case 2 are severely biased, and the biases remain at similar magnitudes even when the

n	PC			Conditional		LBML			PLAC				
			True	Bias	SE	Bias	SE	RE	Bias	SE	SEE	CP	RE
						Case 1: length-biased sampling							
400	50	$\hat{\beta}_1$	1	5	169	−46	115	1.85	10	129	125	94	1.71
		$\hat{\beta}_2$	1	5	150	−49	109	1.60	9	118	113	94	1.60
		$\hat{\Lambda}_{\tau_{30}}$	0.212	1	45	18	40	1.07	−1	41	39	92	1.18
		$\hat{\Lambda}_{\tau_{60}}$	0.546	2	91	42	78	1.04	−1	84	79	93	1.17
	80	$\hat{\beta}_1$	1	24	265	−61	141	2.99	30	169	166	95	2.39
		$\hat{\beta}_2$	1	23	241	−67	133	2.63	30	162	152	94	2.15
		$\hat{\Lambda}_{\tau_{30}}$	0.103	0	37	30	35	0.62	−2	32	30	91	1.31
		$\hat{\Lambda}_{\tau_{60}}$	0.329	−2	92	49	71	1.14	−7	79	76	92	1.37
800	50	$\hat{\beta}_1$	1	8	116	−35	82	1.71	6	89	88	95	1.71
		$\hat{\beta}_2$	1	0	99	−38	75	1.41	4	81	80	94	1.49
		$\hat{\Lambda}_{\tau_{30}}$	0.212	1	30	18	30	0.75	0	28	28	95	1.18
		$\hat{\Lambda}_{\tau_{60}}$	0.546	1	62	36	54	0.90	1	56	56	94	1.22
	80	$\hat{\beta}_1$	1	10	194	−49	101	3.00	14	121	116	94	2.54
		$\hat{\beta}_2$	1	12	203	−48	101	3.31	16	122	116	94	2.73
		$\hat{\Lambda}_{\tau_{30}}$	0.103	−1	31	43	39	0.28	−1	30	30	93	1.08
		$\hat{\Lambda}_{\tau_{60}}$	0.329	−4	62	51	62	0.60	−4	58	58	94	1.15
						Case 2: non-length-biased sampling							
400	50	$\hat{\beta}_1$	1	3	150	−243	103	0.32	3	128	129	94	1.38
		$\hat{\beta}_2$	1	13	157	−232	105	0.38	18	134	129	94	1.36
		$\hat{\Lambda}_{\tau_{30}}$	0.207	−2	40	105	51	0.11	−2	39	38	94	1.02
		$\hat{\Lambda}_{\tau_{60}}$	0.538	−2	65	233	77	0.07	−3	64	64	94	1.01
	80	$\hat{\beta}_1$	1	11	262	−359	117	0.48	27	185	181	95	1.97
		$\hat{\beta}_2$	1	11	260	−364	122	0.46	19	194	181	93	1.78
		$\hat{\Lambda}_{\tau_{30}}$	0.099	−2	34	106	56	0.08	−3	33	31	91	1.04
		$\hat{\Lambda}_{\tau_{60}}$	0.270	−4	61	221	85	0.07	−5	59	58	93	1.05
800	50	$\hat{\beta}_1$	1	−1	107	−227	72	0.20	2	90	91	95	1.44
		$\hat{\beta}_2$	1	2	107	−227	73	0.20	3	92	91	96	1.36
		$\hat{\Lambda}_{\tau_{30}}$	0.207	−1	28	110	40	0.06	−1	27	27	96	1.04
		$\hat{\Lambda}_{\tau_{60}}$	0.538	−1	46	230	56	0.04	−1	45	45	95	1.04
	80	$\hat{\beta}_1$	1	7	176	−347	89	0.24	15	136	130	93	1.66
		$\hat{\beta}_2$	1	−1	179	−348	88	0.25	13	135	129	93	1.75
		$\hat{\Lambda}_{\tau_{30}}$	0.099	−1	25	112	55	0.04	−1	24	23	93	1.02
		$\hat{\Lambda}_{\tau_{60}}$	0.270	−1	44	233	71	0.03	−3	43	42	93	1.02

Table 3.1: Summary of simulation with various sample sizes and censoring rates. PC: censoring percentage; True: true values; Bias, SE, SEE and CP: empirical bias ($\times 10^3$), standard error ($\times 10^3$), standard error estimate ($\times 10^3$) and 95% coverage probability; RE: relative efficiency with respect to the conditional approach estimator (ratio of the mean squared errors). The estimate of $\hat{\Lambda}(t)$ is evaluated at the 30% and 60% percentiles (τ_{30} and τ_{60}) of the observed survival times.

sample size increases to 800.

The proposed method yields considerable efficiency gains compared with the conditional approach estimator under different sample sizes and censoring rates. The efficiency gains in $\hat{\beta}_1$ and $\hat{\beta}_2$ range from 49% to 173% in Case 1 and 36% to 97% in Case 2. The efficiency gains in $\hat{\Lambda}_{\tau_{30}}$ and $\hat{\Lambda}_{\tau_{60}}$ are not as large, but improvement over the conditional approach has been clearly shown, i.e., all relative efficiencies are greater than one. For the length-biased data (Case 1), although the proposed estimator of the regression coefficients has larger standard errors than the LBML estimator, the differences between the two are smaller than the improvement of PLAC estimator achieves over the conditional estimator. Due to smaller biases, the mean squared errors of the PLAC estimator are comparable to those of the LBML estimator. The relative efficiency gains of our estimator increase as the censoring rate increases, because the augmenting pairwise likelihood is not subject to censoring. These higher gains when censoring rate increases are also observed in the LBML estimates, but they are undermined by the simultaneously inflated biases. We also performed additional simulations with the baseline hazard function $\lambda(t) = 1$. The results were as good as those in Table 3.1 or even better with slightly increased efficiency gain and thus are omitted. Taking the biases and the variances altogether, the mean squared errors of our estimator are either the smallest or comparable to the best performer.

Comparing the empirical and estimated standard errors of the proposed estimator, we demonstrate that the variance of the proposed estimator is consistently estimated by the sandwich variance estimator (3.8). We notice that the standard errors for the PLAC estimates under $n = 200$ (Appendix, Table A.1) are approximately twice of those under $n = 800$, which confirms the $\sqrt{n}$-convergence rate as proven in Section

2.3. In the scenarios with $n = 400$ and 80% censoring, the 95% coverage probabilities for the proposed estimator are close to the nominal level, except for $\hat{\Lambda}_{\tau_{30}}$ and $\hat{\Lambda}_{\tau_{60}}$. This is because of the small number of observed events which attenuates the normal approximation not only in our approach, but also in other competitors. For example, the corresponding coverage probabilities of $\hat{\Lambda}_{\tau_{30}}$ using the conditional approach are 91% and 92%, both of which are also below the nominal level. When the sample size increases to 800, all coverage probabilities of the PLAC estimator get closer to the nominal level.

In summary, the proposed estimator performs well under finite sample sizes. It has small empirical biases, and enjoys substantial gains in efficiency in both the regression coefficients and the cumulative baseline hazard function. The performance of our estimator is robust to the violation of the uniform truncation assumption as well as high censoring rates. The proposed sandwich estimator results in good variance estimates for all parameters, and yields reasonable confidence intervals.

3.4 Data Application

We apply the proposed method to the RRI-CKD study introduced in Section 3.1. Investigators were interested in finding the risk of ESRD progression associated with the patient's characteristics at referral. In this study, the survival time was measured from the referral to the composite renal outcome defined as either death, long-term dialysis or kidney transplantation, whichever came first. The truncation time was measured from the referral to the study enrollment. The survival time was also subject to right censoring by non-participating physicians, consent withdraw, lost to follow up, protocol deviation, or the end of study. The baseline patient characteristics included age group (45 to 65 and older than 65), gender, race (white and non-white),

the presence of diabetes, the presence of hypertension, and advanced-stage CKD (defined by estimated GFR less than 30 ml/min/1.73 m^2). Patients without referral information or important covariates were excluded. A total of 545 patients were included in our analysis, of which 256 experienced the composite renal outcome during the study follow-up. The censoring rate was 53%.

We first assessed the uniform truncation assumption using the fact that the observed truncation time A follows the same distribution as the residual survival time V if it holds (Jung, 1999; Mandel and Betensky, 2007). We conducted a paired log-rank test for (A, V), and the null hypothesis of the same distribution was rejected ($p < 0.001$). Moreover, the estimated survival functions for A and V deviate from each other with non-overlapping point-wise confidence intervals (Appendix, Figure A.1). Therefore, we concluded that the uniform truncation assumption did not hold in the data, and hence regression methods for length-biased data might yield invalid inference. The violation of the uniform truncation assumption may be explained by the absence of general guidelines for when to refer to a nephrologist in practice; patients can be referred at either early or late stages of chronic kidney disease.

Table 3.2 gives the regression coefficients estimates and their standard errors from the RRI-CKD data using the conditional approach and the proposed PLAC estimator. Comparing to the conditional approach, we observed consistently smaller standard error estimates and narrower confidence intervals (Appendix, Figure A.2) for all regression coefficients in the analysis of the chronic kidney disease data. The variances ratio of the conditional approach estimate to the corresponding PLAC estimate is 1.30 or greater. This implies that the conditional approach requires at least 30% more CKD patients to achieve the same estimating precision as the PLAC estimator. It is worth noting that the estimated coefficient for Non-White using the

proposed estimator indicates a statistically survival difference between the white and the non-white (estimated hazard ratio is 1.30, $p = 0.045$), whereas the conditional approach estimate does not suggest such a significant difference (estimated hazard ratio is 1.24, $p = 0.185$).

	Conditional			PLAC		
	LHR	SE	p	LHR	SE	p
Older than 65	0.093	0.129	0.473	0.113	0.111	0.311
Male	0.517	0.131	<.001	0.422	0.113	<.001
Non-White	0.213	0.161	0.185	0.262	0.130	0.045
Diabetes	0.424	0.130	0.001	0.507	0.110	<.001
Hypertension	0.168	0.225	0.455	0.075	0.189	0.693
Late-Stage	0.950	0.146	<.001	1.020	0.128	<.001

Table 3.2: Coefficient estimates from the RRI-CKD data using the conditional approach and the proposed method (PLAC). LHR: log hazards ratio (β); SE: standard error; p: p-value.

To illustrate the use of the closed-form variance estimator (3.8), we estimated the survival curves of patients with and without diabetes at referral, and constructed the corresponding 95% point-wise confidence intervals (Figure 3.1). The estimated median survival times and the corresponding confidence intervals, which are the time coordinates of the estimated survival curves and the 95% point-wise confidence intervals crossing the horizontal line at 0.5, are also displayed in Figure 3.1.

3.5 Discussion

We have proposed a semiparametric estimation method for the Cox model with the issue of general left-truncation. By constructing a pairwise likelihood from the marginal likelihood of the truncation times, we have eliminated the unknown truncation distribution from the full likelihood. Based on our simulation studies, the proposed estimator has been shown to be robust to heavy censoring and violation of the uniform truncation assumption, where the robustness means consistency and efficiency gain over the conditional approach estimator across all scenarios considered.

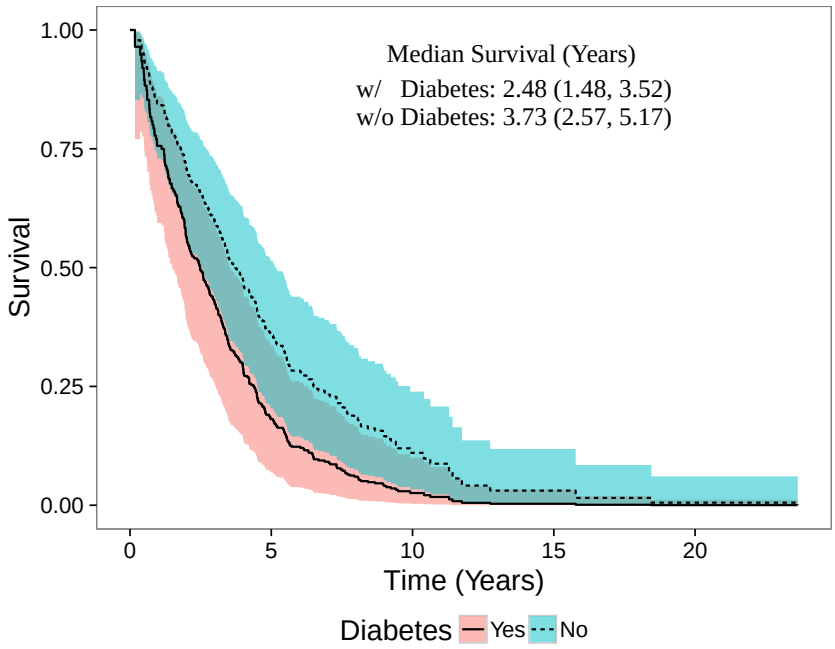

Figure 3.1: Estimated survival curves of patients with diabetes (solid) or without diabetes (dashed) at referral using the proposed method (PLAC). Log-log transformed 95% point-wise confidence intervals are shown as shaded areas. The estimated median survival times for both groups are displayed with the corresponding 95% confidence intervals. The other covariates are set to their reference levels.

On the contrary, length-biased sampling methods of efficiency improvement rely on the uniform truncation assumption to lead to consistent estimates.

We have utilized a nonparametric maximum likelihood approach to estimate the cumulative baseline hazard function along with the regression coefficients. Under regularity conditions, the consistency and asymptotic normality of $(\hat{\beta}, \hat{\Lambda})$ have been rigorously proved which results in a closed-form consistent sandwich variance estimator. We avoid estimating the truncation distribution G, deemed as a nuisance parameter in our application, because eliminating it in the likelihood may simplify inference. The convenience, however, may come at the expense of some efficiency loss. Alternatively, one can estimate G directly and plug the estimate into the full likelihood. Several drawbacks may present. First, if G is estimated nonparametrically, the numerical instability might undermine the estimation of β and Λ. Second, the

inference with plug-in type estimators is challenging. Often the variance estimator is so complicated that resampling-based methods have to be used (Huang et al., 2012). Nevertheless, a full likelihood approach that incorporates additional information in G may yield a more efficient estimator, and further research is warranted.

Even though we relax the uniform assumption in the proposed method, our proposed method, as well as all existing regression methods for length-biased data, still requires independence assumption between A^* and Z^*. Our sensitivity analysis (not shown) indicated the proposed method would yield biased estimates under covariate-dependent truncation. However, under weak dependent cases, we still observed smaller mean squared errors (with 12% biases) compared with the conditional approach estimator. To apply the PLAC estimator, a rigorous model checking tool for the independence assumption between A^* and Z^* is worth pursuing in the future. A graphical inspection tool has been illustrated using the RRI-CKD data in Appendix, Figure A.3 , where we plot and compare the estimated $\hat{G}$ for each level of the covariate under investigation. There is no apparent deviation between the estimated curves for the demographics and hypertension status. As for the CKD stage and diabetes status, the estimated $\hat{G}$ have overlapping confidence intervals. Thus, we conclude that there is no obvious violation of the independence assumption in the RRI-CKD data, which is further supported by the similar point estimates as shown in Table 3.2.

The gain in efficiency is the greatest advantage of the proposed method. For length-biased data, the PLAC estimator is less efficient than the full maximum likelihood estimator of Qin et al. (2011), because the latter is based on the correctly specified uniform truncation distribution. However, the loss of efficiency is not substantial, and the proposed estimator has smaller bias. For non-length-biased data,

if the truncation distribution is known, we can apply the monotone transformation $G(\cdot)$ to both A and X and apply the regression methods for length-biased data on the transformed data as suggested in Huang and Qin (2012). In additional simulations with $A^* \sim \text{Exp}(1)$, we found that the transformation approach out-performed PLAC only when the censoring rate was small to moderate (Appendix, Table A.2). When the censoring rate increased, the transformation approach suffered from large bias and its mean squared errors would be larger than those of the PLAC estimator.

When we combined $\mathcal{L}_n^C$ and $\mathcal{L}_n^P$ in the composite log-likelihood, we used the weights proportional to the reciprocals of their magnitudes (number of terms), which may not be optimal, and further investigation is needed. In the context of additive hazards model, Huang and Qin (2013) studied the optimal weights with which the resulting estimator would have the smallest variance. Their simulation showed that the estimator using the optimal weights was less efficient comparing with the estimator using the reciprocals of the magnitudes as weights. They discussed it was because that the optimal weights involves estimation of the variance of the scores, which requires larger sample sizes to obtain the benefit.

Lastly, while the proposed estimator focuses on handling time-independent covariates, the extension to time-dependent covariates is promising based on our preliminary work. We expect to derive asymptotic properties and devote more effort to reducing computation time, which is magnified by the need of expanding the dataset with the time-dependent covariates.

CHAPTER IV

A Pairwise Likelihood Augmented Cox Estimator with Application to the Kidney Transplantation Registry of Patients under Time-Dependent Treatments

4.1 Introduction

The end stage renal disease (ESRD) is characterized by complete loss of the kidney function filtrating the waste in blood. In the U.S., although the incident rate of ESRD has plateaued after 2001, the prevalence of ESRD continues to rise and reached 1,981 per million on December 31, 2013, an increase of 29% since 2000 (Saran et al., 2015). As the population gets older and the prognosis of the ESRD patients improves, the ESRD prevalence is likely to stay growing in the near future.

Special renal replacement therapy, either dialysis or kidney transplantation, is necessary in order to keep the ERSD patients alive. There are two main modalities of dialysis, hemodialysis (HD) and peritoneal dialysis (PD). On December 31, 2013, 63.9% of the prevalent ESRD cases were treated with HD, 6.9% were on PD, and 29.3% had a functioning kidney transplant (Saran et al., 2015). Kidney transplantation is preferred for long-term survival of the patient (Wolfe et al., 1999). Nevertheless, because of the limited resources, most patients have to be put on dialysis first while waiting for matched organs.

We are interested in the patient survival after the ESRD onset and its association

with the modality of renal replacement therapy (HD, PD or transplantation). To this end, a randomized controlled trial assessing the treatment effect is ideal, yet the only such trial failed due to low patient enrollment and logistic issues (Korevaar et al., 2003). Consequently, investigators usually resort to retrospective cohort study using observational registry data (Vonesh et al., 2006). The Organ Procurement and Transplantation Network (OPTN), operated by the United Network for Organ Sharing (UNOS), contains records on the ESRD patients in U.S. who are listed for kidney transplantation. Because treatment modalities may change over time when ESRD patients receive kidney transplants or resume dialysis following the allografts failures, the treatment modality should be treated as a *time-dependent* covariate (McDonald and Russ, 2002; McDonald and Craig, 2004). Moreover, patients on the OPTN/UNOS waiting list are in general healthier than general ESRD patients, for those who died on dialysis before listed would not be registered (Wolfe et al., 1999). Thus, the survival outcome is also subject to the complication of *left-truncation*. The presence of both time-dependent covariates and left-truncation gives rise to challenges in extending standard approaches to our analysis.

When the survival outcome is left-truncated, survival regression models can be modified to accommodate the biased sampling by conditioning on the truncation times (Kalbfleisch and Lawless, 1991; Wang et al., 1993). The Cox model is a popular semi-parametric model for studying the associations between risk factors and the time-to-event outcome (Cox, 1972). But the conditional approach can be very inefficient when additional distributional assumption can be made on truncation times (Huang and Qin, 2012). For example, if truncation times are uniform, efficiency improvement methods have been proposed using the properties of length-biased sampling (see, e.g., Shen et al., 2016). When the truncation distribution is

independent of the covariates, Huang and Qin (2013) and Wu et al. (2017) propose more efficient estimators for the additive hazards model and the Cox model, respectively. However, all aforementioned authors use time-invariant covariates exclusively in their simulation studies and applications, even though some methods are generalizable to time-dependent covariates. This could be explained by the fact that left-truncation often leaves little covariate information accessible before the study entry, which deters one from considering time-dependent covariates, of which full covariates history is necessary (Huang and Qin, 2012). In the OPTN/UNOS data, the treatment modality at the ESRD onset is less likely to change before listed (enrollment); thus, the full treatment history is available by extrapolating the baseline information on dialysis modality.

In this chapter, an extension to the pairwise likelihood augmented Cox (PLAC) estimator for left-truncated survival data is proposed (cf. Chapter III) to allow for time-dependent covariates. The proposed method eliminates the truncation distribution by a pairwise likelihood argument (Liang and Qin, 2000), and yields more efficient estimator for the regression coefficients and the baseline hazard function compared with the conditional approach under independence assumption between the truncation times and the covariates. A modification of the pairwise likelihood is provided to accommodate the specific dependence between truncation and the covariates met in the OPTN/UNOS registry data. Using empirical process and U-process techniques, the PLAC estimator is shown to be consistent and asymptotically normal with a closed-form variance estimator. Extensive simulation studies demonstrate that the proposed estimator enjoys good finite sample properties and can attain substantial efficiency improvement under various truncation distributions. An iterative algorithm is given, and an R package, `plac`, to implement it is available on CRAN

(R Core Team, 2016).

The rest of this chapter is organized as follows. In Section 4.2, the pairwise likelihood augmented Cox estimator is first derived when the covariates and the truncation time are independent, and then extended to cases where certain dependence exists. Simulation results on the finite sample performance of the proposed estimator are reported in Section 4.3. In Section 4.4, we applied the proposed estimator to the OPTN/UNOS kidney transplant data and compared the results with those from the conditional approach across different states. Discussions of the proposed methods and suggestions for future work are given in Section 4.5. Supplementary materials including the proofs for asymptotic properties and additional simulation and data analysis results are given in Appendix B.

4.2 Proposed Method

4.2.1 Preliminaries

For patients in the *target* population, let T^* be the time from the disease onset to the event of interest, and let $\mathcal{Z}^* \equiv \{\mathbf{Z}^*(t); 0 \leqslant t \leqslant \tau\}$ be the covariate process, where $\mathbf{Z}^*(t)$ is a $p \times 1$ vector of (possibly time-dependent) covariates at t, and τ is a fixed maximum support of follow-up. In the OPTN/UNOS data, T^* is the patient survival time after the ESRD onset, defined using the date of the commencement of renal replacement therapy. Taking HD as the reference modality, $\mathbf{Z}^*(t)$ includes the indicator of PD at time t, the indicator of functioning kidney transplants at time t, and the time-invariant demographics at the baseline. Suppose the Cox proportional hazards model is considered to link the hazard function of T^* to $\mathcal{Z}^*$ (Cox, 1972),

$$\lambda(t \mid \mathcal{Z}^*; \boldsymbol{\beta}) = \lambda(t) \exp\{\boldsymbol{\beta}^T \mathbf{Z}^*(t)\},$$

where $\lambda(\cdot)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients.

In a prevalent cohort, delayed entry may occur for patients who did not participate the study until some time after diagnosis. We define the truncation time, A^*, as the (possibly delayed) study entry time from the onset of disease. Note that 1) T^* and A^* should have the same time origin, and 2) the patients who already had an event of interest (or died) before the study entry (i.e., $T^* < A^*$) are excluded from sampling. Thus, the observed data may consist of a biased sample. To make a clear distinction from the unbiased samples from the target population, we use $\{T, A, \mathcal{Z}\}$ to denote the biased sample of $\{T^*, A^*, \mathcal{Z}^* \mid T^* \geqslant A^*\}$. The biased sampling scheme induces a positive correlation between T and A. When the follow-up time after the study entry (i.e., $T - A$) is subject to potential right censoring, the observed survival time is $X = \min(A + C, T)$, where C is the right-censoring time from the study entry. Let $\Delta = I(T \leqslant A + C)$ be the event indicator, where $I(\cdot)$ denotes the indicator function. We assume C is independent of (T, A) given $\mathcal{Z}$.

Given n independent and identically distributed observations $\{A_i, X_i, \Delta_i, \mathcal{Z}_i; i = 1, \ldots, n\}$ from a prevalent cohort, the full likelihood under the Cox model is proportional to

$$\mathcal{L}_n = \prod_{i=1}^{n} \frac{f(X_i \mid \mathcal{Z}_i)^{\Delta_i} S(X_i \mid \mathcal{Z}_i)^{1-\Delta_i} dG(A_i)}{\int S(a \mid \mathcal{Z}_i) dG(a)},$$

where the density and survival functions of T^* given the covariate process are denoted by $f(\cdot \mid \mathcal{Z})$ and $S(\cdot \mid \mathcal{Z})$, the unspecified marginal distribution function of A^* is denoted by $G(\cdot)$, and the integrals without the domain of integration are taken over the entire follow-up period $[0, \tau]$. The full likelihood can be decomposed into two

parts:

$$\mathcal{L}_n = \prod_{i=1}^{n} \frac{f(X_i \mid \mathcal{Z}_i)^{\Delta_i} S(X_i \mid \mathcal{Z}_i)^{1-\Delta_i}}{S(A_i \mid \mathcal{Z}_i)} \times \prod_{i=1}^{n} \frac{S(A_i \mid \mathcal{Z}_i) dG(A_i)}{\int S(a \mid \mathcal{Z}_i) dG(a)} \equiv \mathcal{L}_n^C \times \mathcal{L}_n^M, \tag{4.1}$$

where $\mathcal{L}_n^C$ is the conditional likelihood of (X, Δ) given $(A, \mathcal{Z})$, and $\mathcal{L}_n^M$ is the marginal likelihood of A given $\mathcal{Z}$. The conditional approaches use $\mathcal{L}_n^C$ for estimating the parameters of the Cox model, and is shown to be fully efficient if no further distributional assumption can be made on A^* (Kalbfleisch and Lawless, 1991; Wang et al., 1993). However, in practice, if additional assumption holds for the truncation times A^*, we can exploit $\mathcal{L}_n^M$ to obtain more precise estimates.

Most literature requires G to be independent of $\mathcal{Z}$, and that it is either fully specified (e.g., uniform as in length-biased sampling regression methods) or has a parametric form (Shen et al., 2016; Liu et al., 2016). These parametric assumptions are sometimes too strong and subject to model misspecification. In fact, the independence between $\mathcal{Z}$ and the underlying truncation time A^* is adequate to improve the estimation efficiency (Huang and Qin, 2013; Wu et al., 2017). In Section 4.2.2, the derivation of the pairwise likelihood relies on this assumption to eliminate the unspecified distribution function G. When dealing with time-dependent covariates, this independence assumption is often violated. However, in Section 4.2.3, we will show the pairwise likelihood can be modified to retain the same form under certain types of dependence between $\mathcal{Z}$ and A^*, such as the time-dependent modality in the OPTN/UNOS data.

4.2.2 The PLAC Estimator for Data with Time-Dependent Covariates

Suppose $\mathcal{Z}$ and the *underlying* truncation time A^* are independent, i.e., $\mathcal{Z} \perp A^*$ or $G(\cdot \mid \mathcal{Z}) = G(\cdot)$. We first derive a pairwise likelihood $\mathcal{L}_n^P$ from $\mathcal{L}_n^M$ such that

retains the major marginal information depending only on $(\boldsymbol{\beta}, \Lambda)$. Let

$$\mathcal{L}_{i|j}^{M} = S(A_i \mid \mathcal{Z}_j) dG(A_i) \{ \int S(a \mid \mathcal{Z}_j) dG(a) \}^{-1}. \tag{4.2}$$

Conditional on the order statistics of (A_i, A_j) and $(\mathcal{Z}_i, \mathcal{Z}_j)$, the pseudo-likelihood of the pair (i, j), $i < j$, is

$$\mathcal{L}_{ij}^{P} = \frac{\mathcal{L}_{i|i}^{M} \cdot \mathcal{L}_{j|j}^{M}}{\mathcal{L}_{i|i}^{M} \cdot \mathcal{L}_{j|j}^{M} + \mathcal{L}_{i|j}^{M} \cdot \mathcal{L}_{j|i}^{M}} = \left\{ 1 + \frac{S(A_i \mid \mathcal{Z}_j) S(A_j \mid \mathcal{Z}_i)}{S(A_i \mid \mathcal{Z}_i) S(A_j \mid \mathcal{Z}_j)} \right\}^{-1},$$

where the second and third factors in (4.2), depending solely on A and $\mathcal{Z}$, respectively, are eliminated by $\mathcal{Z} \perp A^*$. Multiplying over all ordered pairs, the pairwise likelihood $\mathcal{L}_n^P = \prod_{i<j} (1 + R_{ij}(\boldsymbol{\beta}, \Lambda))^{-1}$, where the generalized odds ratio (GOR; Liang and Qin, 2000)

$$R_{ij}(\boldsymbol{\beta}, \Lambda) = \exp \left[\int (e^{\boldsymbol{\beta}^T \mathbf{Z}_i(t)} - e^{\boldsymbol{\beta}^T \mathbf{Z}_j(t)}) \{ I(t \leqslant A_i) - I(t \leqslant A_j) \} d\Lambda(t) \right]. \tag{4.3}$$

We augment $\mathcal{L}_n^C$ with $\mathcal{L}_n^P$, which yields the composite log-likelihood $n^{-1} \log \mathcal{L}_n^C(\boldsymbol{\beta}, \Lambda) + 2\{n(n-1)\}^{-1} \log \mathcal{L}_n^P(\boldsymbol{\beta}, \Lambda)$. Note that we normalize $\log \mathcal{L}_n^C$ and $\log \mathcal{L}_n^P$ by different factors in order to account for their different magnitudes; there are n terms in $\log \mathcal{L}_n^C$ and $2^{-1} n(n-1)$ terms in $\log \mathcal{L}_n^P$.

Assume $\Lambda(0) = 0$, and that $\Lambda(\cdot)$ is a step function with positive jumps $\Lambda\{\cdot\}$ only at the observed event times $X_i \mid \Delta_i = 1$. Let $w_1 < \cdots < w_m$ $(m \leqslant n)$ be the ordered distinct observed event times, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^T$, where $\lambda_k = \Lambda\{w_k\}$. For $k = 0, 1$ and 2, define the following functions which appear in $\log \mathcal{L}_n^P$ and its derivatives:

$$Q_{ij}^{(k)}(t; \boldsymbol{\beta}) = \left(\mathbf{Z}_i(t)^{\otimes k} e^{\boldsymbol{\beta}^T \mathbf{Z}_i(t)} - \mathbf{Z}_j(t)^{\otimes k} e^{\boldsymbol{\beta}^T \mathbf{Z}_j(t)} \right) (I(t \leqslant A_i) - I(t \leqslant A_j)),$$

where, for $a \in \mathbb{R}^p$, $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$. We will use R_{ij} and $Q_{ij}^{(k)}(t)$ below in stead of $R_{ij}(\boldsymbol{\beta}, \Lambda)$ and $Q_{ij}^{(k)}(t; \boldsymbol{\beta})$ when the meaning of the notations is clear

from the context. Replacing $\lambda(t)$ with $\Lambda\{t\}$ in the composite log-likelihood, we have

$$\ell_n^c(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left[\Delta_i\left\{\log\Lambda\{X_i\} + \boldsymbol{\beta}^T\mathbf{Z}_i(X_i)\right\} - \sum_{k=1}^{m} Y_i(w_k)\exp\{\boldsymbol{\beta}^T\mathbf{Z}_i(w_k)\}\lambda_k\right]$$

$$- \frac{2}{n(n-1)}\sum_{i<j}\log\{1 + R_{ij}(\boldsymbol{\beta}, \boldsymbol{\lambda})\}, \tag{4.4}$$

where $R_{ij}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \exp\{\sum_{k=1}^{m}\lambda_k Q_{ij}^{(0)}(w_k)\}$, and $Y_i(t) = I(A_i \leqslant t \leqslant X_i)$. We maximize (4.4) as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and refer to the maximizer $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ (or $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$) as the pairwise likelihood augmented Cox (PLAC) estimator, where $\hat{\Lambda}(t) = \sum_{k=1}^{m}\hat{\lambda}_k I(w_k \leqslant t), t \in [0, \tau]$.

Differentiating (4.4) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ yields the composite score functions:

$$U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\Delta_i\mathbf{Z}_i(X_i) - \sum_{k=1}^{m}\lambda_k Y_i(w_k)\mathbf{Z}_i(w_k)e^{\boldsymbol{\beta}^T\mathbf{Z}_i(w_k)}\right\}$$
$$- \frac{1}{n(n-1)}\sum_{i\neq j}\frac{\sum_{k=1}^{m}\lambda_k Q_{ij}^{(1)}(w_k)}{1 + R_{ij}^{-1}},$$
$$U_{\lambda_k}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = w_k)\left\{\frac{\Delta_i}{\lambda_k} - Y_i(w_k)e^{\boldsymbol{\beta}^T\mathbf{Z}_i(w_k)}\right\}$$
$$- \frac{1}{n(n-1)}\sum_{i\neq j}\frac{Q_{ij}^{(0)}(w_k)}{1 + R_{ij}^{-1}}.$$

Let $U_{\boldsymbol{\lambda}}^T = (U_{\lambda_1}, \ldots, U_{\lambda_m})$, then $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ solves $U(\boldsymbol{\beta}, \boldsymbol{\lambda}) \equiv (U_{\boldsymbol{\beta}}^T, U_{\boldsymbol{\lambda}}^T)^T(\boldsymbol{\beta}, \boldsymbol{\lambda}) = 0$. Directly solving these nonlinear equations with complicated pairwise structure is difficult, since we can no longer profile out Λ as in the conditional approach to get the partial likelihood (Cox, 1975). Therefore, we propose the following algorithm to solve for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\lambda}}$ iteratively:

Step 1. Start with initial values $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\lambda}^{(0)}$.

Step 2. At the r-th iteration, update each $\lambda_k^{(r)}$ using

$$\lambda_k^{(r)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\Delta_i I(X_i = w_k)}{\frac{1}{n}\sum_{i=1}^{n} Y_i(w_k)e^{\mathbf{Z}_i(w_k)^T\boldsymbol{\beta}^{(r-1)}} + \frac{1}{n(n-1)}\sum_{i\neq j}\frac{Q_{ij}^{(0)}(w_k;\boldsymbol{\beta}^{(r-1)})}{1+1/R_{ij}(\boldsymbol{\beta}^{(r-1)},\boldsymbol{\lambda}^{(r-1)})}}. \tag{4.5}$$

Step 3. Update $\boldsymbol{\beta}^{(r)}$ by one step of Newton-Raphson iteration:

$$\boldsymbol{\beta}^{(r)} = \boldsymbol{\beta}^{(r-1)} - \left\{ \dot{U}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}^{(r)}) \right\}^{-1} \left\{ U_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}^{(r)}) \right\},$$

where $\dot{U}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}^{(r)}) = \partial U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) / \partial \boldsymbol{\beta}^T \big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(r-1)}, \boldsymbol{\lambda}=\boldsymbol{\lambda}^{(r)}}$.

Step 4. Repeat Step 2 and 3 until the algorithm converges.

The estimates from the conditional approach can be used to set the initial values; our simulations demonstrated that the algorithm is robust to the choice of initial values. Moreover, the self-consistent solution (4.5) alleviates the vast computation burden of optimizing over the infinite-dimensional parameter Λ, which makes the algorithm manageable even on personal computers.

4.2.3 The Modified Pairwise Likelihood

For a patient, let ζ denote the time from the ESRD onset to accepting a kidney transplant. By the nature of the waiting list, the OPTN/UNOS data only consists of the patients who get transplanted *after* listed (including the pre-emptive transplantations), and those who are never transplanted and are still waiting for a kidney (i.e., the transplant happened hypothetically after their failure or censoring events). This indicates that in the observed sample, $\zeta > A^*$ for all subjects. However, as illustrated in Figure 4.1, the independence between the covariates and the underlying truncation time actually assumes the transplant status can also change *before* listed. Let $\zeta = A^* + \zeta_w$, then equivalently $\zeta_w > 0$ for everyone in the sample. Due to the constraints on either ζ or ζ_w, dependence exists between $\mathcal{Z}$ and A^* in the OPTN/UNOS data. In our simulation (not shown), we found that using $\mathcal{L}_n^P$ derived in Section 4.2.2 naïvely could lead to biased estimates because of the violation of the independence assumption between $\mathcal{Z}$ and A^*.

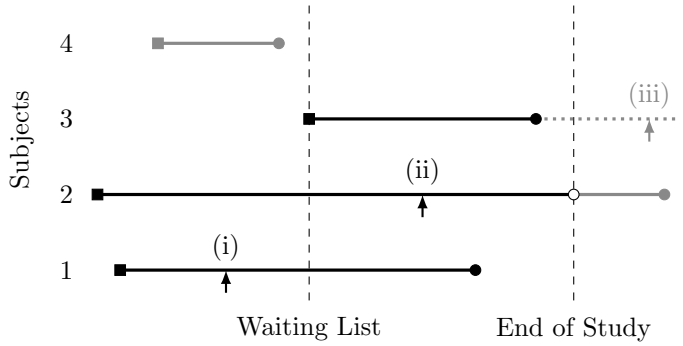

Figure 4.1: Examples of different follow-up scenarios in left-truncated right-censored data.

■ : onset; ● : event; ○ : censored. ↑ : transplant status change. (i): $\zeta < A^* < T^* < A^* + C$; (ii): $A^* < \zeta < A^* + C < T^*$; (iii): $A^* < T^* < \zeta$; Subject 4 is left-truncated ($T^* < A^*$).

Without loss of generality, let $\mathcal{Z} = \{(Z_f, Z_v(t)), 0 \leqslant t \leqslant \tau\}$, where the second covariate is time-dependent, and let $\boldsymbol{\beta}$ be partitioned as (β_f, β_v) correspondingly. We consider the time-dependent covariates which depend on A^* taking the form

$$Z_v(t; A^*) = Z_v^*(t)I(t \leqslant A^*) + Z_v^\dagger(t)I(t > A^*),$$

where $\{Z_v^*(t), 0 \leqslant \tau\}$ and $\{Z_v^\dagger(t), 0 \leqslant \tau\}$ are the pre- and post-truncation covariate processes corresponding to $Z_v(t)$. For example, the transplant indicator $Z_v(t) = I(t \geqslant \zeta)$ has $Z_v^*(t) \equiv 0$ and $Z_v^\dagger(t) = I(t \geqslant \zeta)$. We use the notation $\mathcal{Z}(a^*)$ to emphasize its dependence on the *underlying* truncation time a^*. For a subject from the prevalent cohort, consider the marginal likelihood $\mathcal{L}^M = S(A \mid \mathcal{Z}(A))dG(A)\{\int S(a \mid \mathcal{Z}(a))dG(a)\}^{-1}$. For any given a, since $I(t > a) = 0$ if $t \leqslant a$, the numerator and the denominator of $\mathcal{L}^M$ are given by

$$\exp\left[-\int I(t \leqslant A)e^{\beta_f^T Z_f + \beta_v\{Z_v^*(t)I(t\leqslant A) + Z_v^\dagger(t)I(t>A)\}}d\Lambda(t)\right]dG(A) \tag{4.6}$$

$$= \exp\left\{-\int I(t \leqslant A)e^{\beta_f^T Z_f + \beta_v Z_v^*(t)}d\Lambda(t)\right\}dG(A)$$

$$\text{and} \quad \int \exp\left[-\int I(t \leqslant a)e^{\beta_f^T Z_f + \beta_v\{Z_v^*(t)I(t\leqslant a) + Z_v^\dagger(t)I(t>a)\}}d\Lambda(t)\right]dG(a) \tag{4.7}$$

$$= \int \exp\left\{-\int I(t \leqslant a)e^{\beta_f^T Z_f + \beta_v Z_v^*(t)}d\Lambda(t)\right\}dG(a).$$

It is worth noting that $Z_v(t;a)$ in (4.7) depends on a in stead of the *observed* A of the subject as in (4.6). Recall that (4.7) is the probability of not being truncated ($T^* \geqslant A^*$) for subject with characteristics $\mathcal{Z}$. The dependence of $\mathcal{Z}$ on A^* motivates us to treat the *characteristics* of the subject as a process $\{\mathcal{Z}(a); 0 \leqslant a \leqslant \tau\}$, instead of the observed $\mathcal{Z}(A)$ only, which results in the modified pairwise likelihood $\mathcal{L}_n^{P*} = \prod_{i<j}(1 + R_{ij}^*(\boldsymbol{\beta}, \Lambda))^{-1}$, where

$$R_{ij}^*(\boldsymbol{\beta}, \Lambda) = \exp\left[\int \{e^{\boldsymbol{\beta}^T \mathbf{Z}_i^*(t)} - e^{\boldsymbol{\beta}^T \mathbf{Z}_j^*(t)}\}\{I(t \leqslant A_i) - I(t \leqslant A_j)\} d\Lambda(t)\right], \tag{4.8}$$

where $\mathbf{Z}_i^*(t) = (Z_{fi}, Z_{vi}^*(t))^T$ and $\mathbf{Z}_j^*(t) = (Z_{fj}, Z_{vj}^*(t))^T$. It is worth noting that the time-dependent covariates in the modified $\mathcal{L}_n^{P*}$ depends only on the related *pre-truncation* processes, even if we *observe* the *post-truncation* processes after the subject is enrolled. It also indicates that we need to extrapolate the pre-truncation processes beyond the observed truncation times to obtain the modified pairwise likelihood. In our simulation, we find that as long as the processes $\{Z_{vi}^*(t), 0 \leqslant \tau\}, i = 1, \ldots, n$, are available and vary between subjects, the efficiency of β_v can be improved using the PLAC estimator. In the example of transplant indicator, however, by (4.8), the GOR for the transplant indicator $Z_v(t) = I(t \geqslant A + \zeta_w)$ simplifies to $R_{ij}^*(\boldsymbol{\beta}, \Lambda) = \exp\left[(e^{\beta Z_{fi}} - e^{\beta Z_{fj}})\{\Lambda(A_i) - \Lambda(A_j)\}\right]$. It means that when all subjects $Z_{vi}^*(t) \equiv Z_v^*(t)$, $i = 1, \ldots, n$, no information can be gained from $\mathcal{L}_n^{P*}$ for estimating β_v.

4.3 Simulation

We investigated the finite-sample performance of the proposed method through extensive simulation studies and compared it with the conditional approach. In the first set of simulations, the underlying truncation time A^* was generated from the $U(0, 100)$, and the underlying survival time T^* from a Cox model with two

independent covariates: $\lambda(t|\mathcal{Z}) = \lambda(t)\exp(\beta_f Z_f + \beta_v Z_v(t))$, where $\beta_f = \beta_v = 1$ and $\lambda(t) = 2t$. The time-invariant covariate $Z_f \sim U(-1, 1)$. For the time-dependent covariate $Z_v(t)$, we considered three cases: (1) $Z_v(t) = I(t \geqslant \zeta)$; (2) $Z_v(t) = I(t \geqslant A^* + \zeta_w)$; (3) $Z_v(t) = \eta I(t \leqslant A^* + \zeta_w) - I(t > A^* + \zeta_w)$, where ζ and $\zeta_w \sim$ Exponential(1), and $\eta \sim$ Bernoulli(0.5). Case 1 represents the scenario in which $\mathcal{Z} \perp A^*$, whereas Case 2 and 3 are examples of $\mathcal{Z} \not\perp A^*$. The transplant status is of Case 2; it can only change to one after enrollment (listed). Case 3 resembles the time-dependent treatment in the OPTN/UNOS data, where different dialysis modalities could be used before enrollment. Censoring time was generated from $U(0, C_{\max})$, where $C_{\max}$ was chosen to designate different censoring rates.

In the second set of simulations, we generated data under Case 1 and varied the underlying truncation distribution G to investigate its impact on the performance of the proposed method. Specifically, we consider three continuous distributions: exponential with rate 1 (Exp(1)), Weibull distribution with shape and scale parameters equal 3 (WB(3, 3)) and U(0, 100) to represent the cases where the patients entering the study at the earlier stages, later stages or uniformly. Correspondingly, we also consider three discrete distributions: binomial with 5 trials and success probability of 0.2 and 0.8 (Bin(5, .2) and Bin(5, .8)) and discrete uniform distribution on integers from 0 to 5 (DU(0:5)).

Under all scenarios, 1000 datasets with $N = 400$ were generated. We estimated β_f, β_v, and $\Lambda(t)$ at fixed time points, τ_{30} and τ_{70}, the 30% and 70% percentiles of the observed survival times. Table 4.1 and 4.2 summarize simulation results under the two settings, respectively, reporting the average difference of the estimate from the true value (Bias), the empirical standard error of the estimate (SE), the average standard error estimate (SEE), the 95% coverage probability (CP), and the relative

efficiency (RE), which is the ratio of mean squared errors between the conditional approach and the PLAC estimators.

PC		True	Conditional		PLAC				
			Bias	SE	Bias	SE	SEE	CP	RE
					Case 1: $Z_v(t) = I(t \geqslant \zeta)$				
0	β_f	1	4	105	4	92	90	94.4	1.30
	β_v	1	−1	108	0	97	93	93.8	1.22
	$\Lambda_0(\tau_{30})$	0.48	−1	56	−2	55	53	94.1	1.05
	$\Lambda_0(\tau_{70})$	1.38	3	128	1	124	115	93.2	1.07
50	β_f	1	9	150	11	121	114	92.7	1.52
	β_v	1	17	154	16	128	122	94.6	1.44
	$\Lambda_0(\tau_{30})$	0.28	−2	52	−3	49	45	91.5	1.09
	$\Lambda_0(\tau_{70})$	0.86	−7	109	−8	104	100	93.5	1.09
80	β_f	1	7	233	23	159	151	94.8	2.11
	β_v	1	−2	245	10	172	163	94.6	2.02
	$\Lambda_0(\tau_{30})$	0.14	−2	44	−4	42	39	90.4	1.11
	$\Lambda_0(\tau_{70})$	0.57	1	116	−3	107	106	94.5	1.18
					Case 2: $Z_v(t) = I(t \geqslant A^* + \zeta_w)$				
0	β_f	1	5	107	6	94	90	95.8	1.31
	β_v	1	2	107	0	106	103	94.8	1.02
	$\Lambda_0(\tau_{30})$	0.62	1	70	0	68	65	93.0	1.03
	$\Lambda_0(\tau_{70})$	1.89	3	148	3	143	138	93.6	1.07
50	β_f	1	0	144	7	117	113	93.5	1.51
	β_v	1	0	159	−3	157	151	95.0	1.02
	$\Lambda_0(\tau_{30})$	0.37	0	61	−1	60	58	93.4	1.05
	$\Lambda_0(\tau_{70})$	1.26	5	132	5	129	129	94.4	1.05
80	β_f	1	0	239	23	166	154	93.0	2.04
	β_v	1	−10	319	−19	315	293	94.2	1.03
	$\Lambda_0(\tau_{30})$	0.20	0	59	−1	57	53	90.2	1.04
	$\Lambda_0(\tau_{70})$	0.90	3	160	1	153	146	94.2	1.10
					Case 3: $Z_v(t) = \eta I(t \leqslant A^* + \zeta_w) - I(t > A^* + \zeta_w)$				
0	β_f	1	4	101	4	92	91	94.5	1.19
	β_v	1	5	83	4	78	77	94.4	1.12
	$\Lambda_0(\tau_{30})$	0.58	0	55	0	55	54	94.2	1.01
	$\Lambda_0(\tau_{70})$	2.48	9	187	8	184	177	93.9	1.03
50	β_f	1	−1	134	4	115	118	95.6	1.35
	β_v	1	0	118	1	107	103	93.8	1.22
	$\Lambda_0(\tau_{30})$	0.33	2	46	2	45	45	93.7	1.03
	$\Lambda_0(\tau_{70})$	1.22	8	119	8	119	120	95.7	1.02
80	β_f	1	6	240	22	172	166	94.1	1.92
	β_v	1	−2	233	9	177	169	94.4	1.72
	$\Lambda_0(\tau_{30})$	0.15	2	41	0	38	36	92.1	1.15
	$\Lambda_0(\tau_{70})$	0.66	3	137	4	132	126	92.1	1.08

Table 4.1: Summary of simulation with various cases for $Z_v(t)$. PC: censoring rate. True: true parameter value. Bias and SE: empirical bias (×1000) and standard error (×1000); SEE: estimated standard error (×1000); CP: 95% coverage probability; RE: relative efficiency. τ_{30} and τ_{70} are the 30% and 70% quantiles of observed event times.

Table 4.1 demonstrates that the proposed method yields parameter estimates close

to the true values in all cases for $Z_v(t)$. Using the PLAC estimator, all parameters estimates have smaller empirical standard errors compared with those of the conditional approach. It worth noting that the gain in efficiency is larger under higher censoring rates; under 80% of censoring, the precision of the PLAC estimator for the β's is twice better. Even with no censoring, efficiency improvement of 10% to 30% can still be observed. All REs in Case 2 for β_v are close to one, which confirms the remark in Section 4.2.3 about no extra information in $\mathcal{L}_n^P$ for estimating β_v. On the other hand, $Z_v(t)$ in Case 3 varies across subjects before enrollment, which provides additional pre-truncation information about β_v, so that better estimation precision for it is also obtained. Lastly, the sandwich standard error estimates are close to their empirical counterparts, and the corresponding coverage probabilities are close to the nominal level.

Similarly, in Table 4.2, the PLAC estimator is consistent and more efficient under all truncation distributions considered. Recalling no censoring was assumed here, we would expect more efficiency gain (i.e., higher RE) in practice when a higher censoring presents. The improvement in estimation efficiency for β's ranges from 13% to 63%, whereas that for $\Lambda_0(\tau_{30})$ and $\Lambda_0(\tau_{70})$ is between 3% and 60%. Better efficiency gains are obtained when truncation occurs at later times, which is consistent with the findings in Huang and Qin (2013). The slightly below-nominal coverage for $\Lambda_0(\tau_{30})$ when G is Weibull is due to the lack of observed events at earlier times.

We also considered different G for the $\mathcal{Z} \not\perp A^*$ cases, and the results were similar to those in Tables 4.1 and 4.2. In addition, when we increase the sample size, the decrease of the empirical standard error for all parameters approximate the $\sqrt{n}$ convergence rate, as proved in Appendix B. Finally, we checked the sensitivity of the proposed methods to various types of baseline hazard functions and $Z_v(t)$, and the

results were either better than or similar to those reported here. The detailed simulation settings and results for the sensitivity analysis are also provided in Appendix B.

G		True	Conditional		PLAC				
			Bias	SE	Bias	SE	SEE	CP	RE
Exp(1)	β_f	1	4	102	3	94	92	93.9	1.17
	β_v	1	−1	106	−1	100	97	94.4	1.13
	$\Lambda_0(\tau_{30})$	0.40	1	44	0	43	44	94.2	1.03
	$\Lambda_0(\tau_{70})$	1.15	6	98	5	95	98	95.2	1.05
U(0, 100)	β_f	1	1	101	2	91	90	94.7	1.25
	β_v	1	−2	107	−5	95	93	94.7	1.26
	$\Lambda_0(\tau_{30})$	0.48	3	55	2	54	53	94.4	1.04
	$\Lambda_0(\tau_{70})$	1.38	10	124	10	119	116	94.5	1.08
WB(3, 3)	β_f	1	8	123	7	96	90	92.6	1.63
	β_v	1	−1	109	1	89	86	93.8	1.48
	$\Lambda_0(\tau_{30})$	1.16	−15	193	−25	169	138	89.2	1.29
	$\Lambda_0(\tau_{70})$	2.68	−4	276	−18	246	221	92.8	1.25
Bin(5, .2)	β_f	1	9	102	7	90	90	94.1	1.29
	β_v	1	6	105	6	97	95	94.0	1.18
	$\Lambda_0(\tau_{30})$	0.32	−1	34	−1	33	34	94.9	1.04
	$\Lambda_0(\tau_{70})$	1.27	1	114	−1	109	108	94.7	1.09
DU(0:5)	β_f	1	5	106	7	94	90	93.4	1.25
	β_v	1	5	104	4	95	96	94.7	1.20
	$\Lambda_0(\tau_{30})$	0.30	−1	32	−1	32	32	94.5	1.03
	$\Lambda_0(\tau_{70})$	1.21	4	110	3	105	104	94.0	1.11
Bin(5, .8)	β_f	1	3	125	8	99	96	94.2	1.58
	β_v	1	4	106	5	86	86	95.3	1.53
	$\Lambda_0(\tau_{30})$	1.31	5	216	−7	170	164	94.4	1.61
	$\Lambda_0(\tau_{70})$	3.80	14	439	1	362	348	93.9	1.47

Table 4.2: Summary of simulation with various G under Case 1 with no censoring. True: true parameter value. Bias and SE: empirical bias (×1000) and standard error (×1000); SEE: estimated standard error (×1000); CP: 95% coverage probability; RE: relative efficiency. τ_{30} and τ_{70} are the 30% and 70% quantiles of observed event times.

4.4 Data Application

We applied the proposed method to the OPTN/UNOS kidney transplant registry data introduced in Section 4.1 and compared the results with those from the conditional approach. Beside the time-dependent treatment for ESRD patients, the model also adjusted for the age at disease onset, ethnicity and gender. The survival outcome of interest was defined as the time from the commencement of first renal replacement therapy (regularly administered dialysis or kidney transplantation) for ESRD

to death. Our analysis was as-treated, and for simplicity, patients whose first allograft failed and went back on dialysis or listed for re-transplantation thereafter were censored at the 90th day following the graft failure, unless the patient died within this 90-day window. A similar rule was used in McDonald and Craig (2004) to avoid inferential results in favor of the transplantation, because otherwise the death right after the allograft failure will be attributed to the subsequent dialysis. Note that the dialysis patients who died before entering the waiting list were left truncated, and the UNOS/OPTN data contain no information about them. In addition to the left truncation, the event times were also subject to the right censoring by dropout, 90 days after after graft failure and the resumption of maintenance dialysis or the end of the follow-up in December, 2015.

Our analysis included the ESRD patients who started their renal replacement therapy later than January 1st, 1995 from 40 states in the US and were listed for their first kidney transplantation in the year 2006. Most patients started with regularly administered dialysis, either HD or PD, yet we also included those who underwent pre-emptive transplantations; they did not go through any dialysis before transplantation. Although most patients on the waiting list initiated their renal replacement therapy before the enrollment, incidence cases were still possible since ESRD patients could be listed because of low glomerular filtration rate, and that we observed their commencement of the treatment during the follow-up. Sample sizes, number of deaths, and censoring rates of the included states are listed in Appendix, Table B.5. Except those who died in the 90-day window after graft failure, death was attributed to the treatment modality of the patient was taking at their time of death. The reference group for the treatment was HD. For each state, Cox model was assumed and the coefficient were estimated using both the conditional approach (Conditional)

and the proposed method (PLAC).

Figure 4.2 provide a Christmas tree plot for the PD and TX coefficient estimates from all the included states. Most of the PD coefficient estimates are positive except those in Pennsylvania, Tennessee, Arizona, Arkansas and Kansas, though few are statistically significant. This indicates receiving PD while waiting for transplantation is possibly associated with worse long-term survival for ESRD patients compared to undergoing HD. The negative TX coefficient estimates, on the other hand, confirms the better prognosis for transplanted ESRD patients. The magnitudes of the beneficial effect, however, also vary a lot across different states. Compared with the conditional approach, the proposed method give similar point estimates for both coefficients. It also yields shorter confidence intervals for PD coefficient in most states, whereas the confidence intervals for the TX coefficient are usually slightly wider. The maps in Figure 4.3 show the magnitudes of the estimated hazards ratio for PD or TX over HD using the PLAC estimator. The south-east states seems to have more adverse effect on patient survival for PD, whereas the geographical pattern for the beneficial effect of TX are not clear from the analysis.

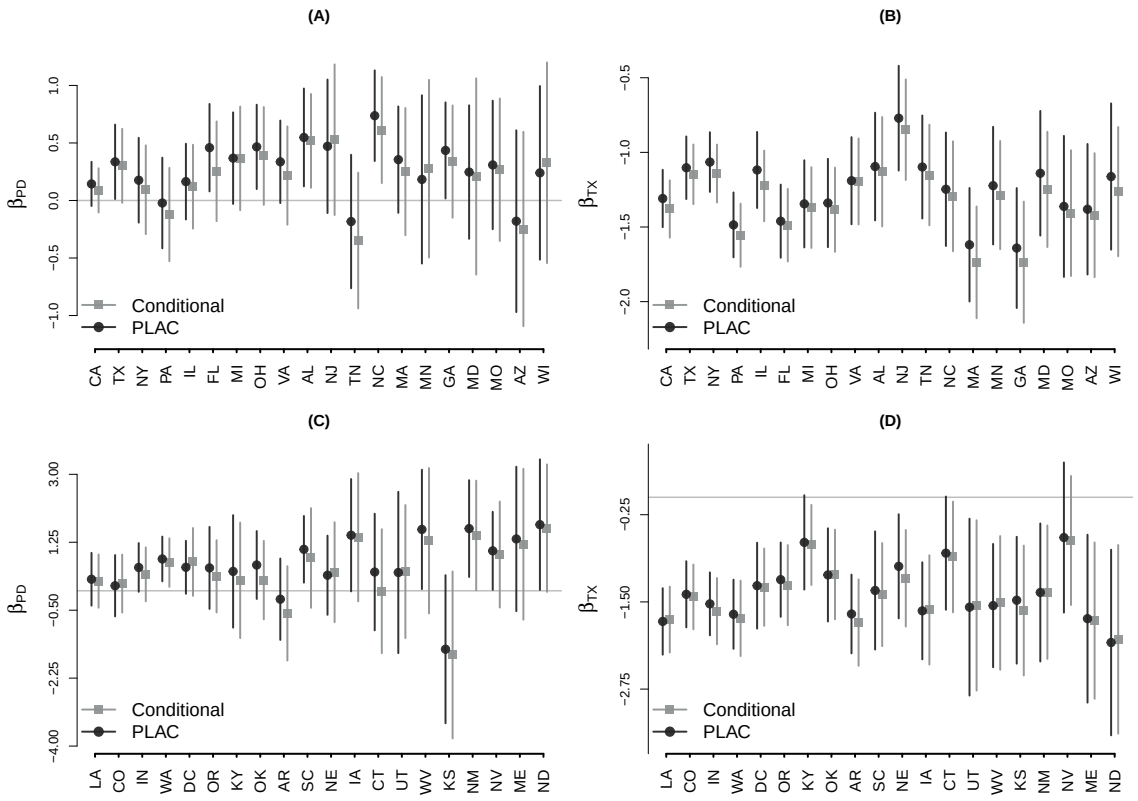

Figure 4.2: Christmas tree plot for the log hazard ratio estimates for PD and TX for the included states. The four panels are respectively (A) β_{PD} estimates for large states; (B) β_{TX} estimates for large states; (C) β_{PD} estimates for small states; (D) β_{TX} estimates for small states. The black round dots and the grey square dots represent the point estimates for the PLAC estimator and the conditional approach estimator, respectively. The verticle lines with the same color give the 95% confidence intervals.

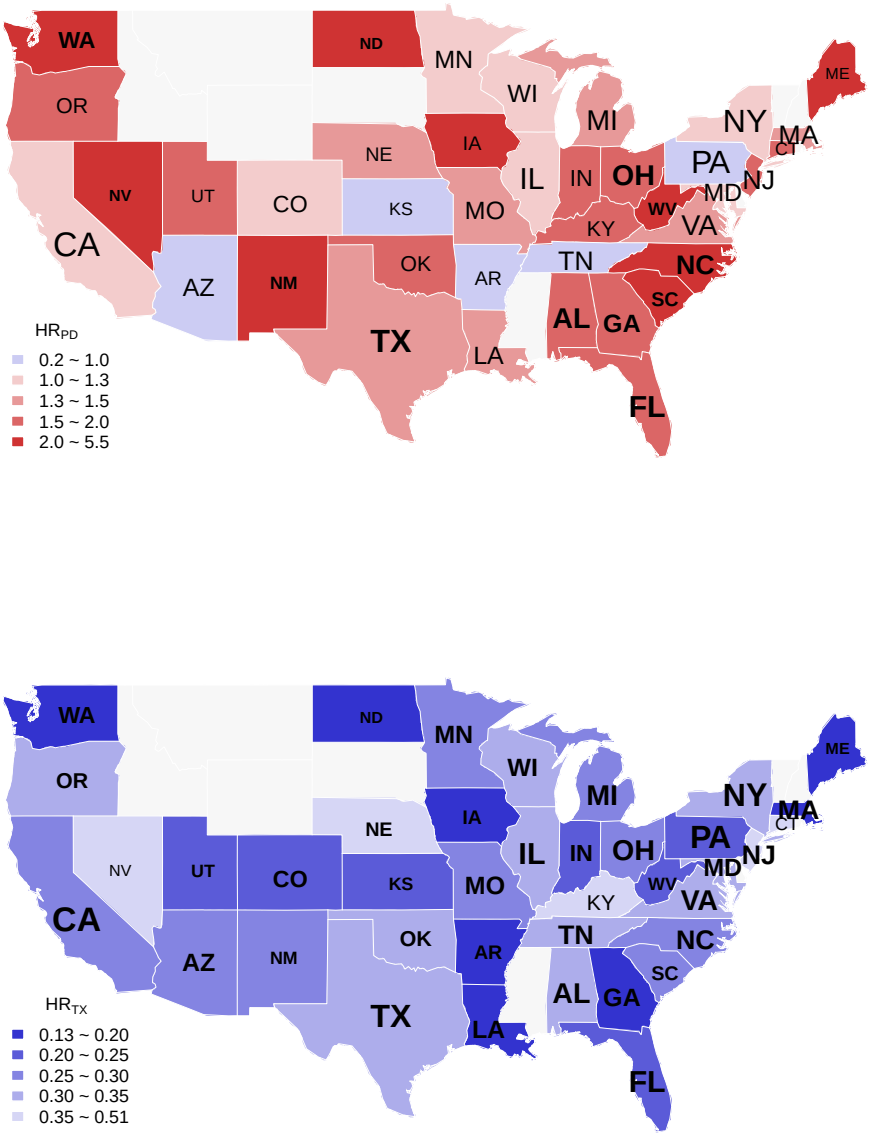

Figure 4.3: US maps of the hazards ratio estimates for PD and TX to HD using the proposed method. Different colors indicate different magnitudes. States in blue means the corresponding treatment is beneficial compared with the reference, whereas stats in red implies the treatment increases of deaths in long term. The darker the color, the further away the effect is from the null, and the statistically significant ones has the states abbreviations in bold.

Table 4.3 gives the summary of the model fitting results using both the PLAC and the conditional approach with the OPTN/UNOS data in Ohio and West Virginia.

These two states were chosen to show the efficiency improvement in the demographics, where all related coefficients are estimated with smaller standard errors using the proposed PLAC estimator. Note the PLAC estimate of transplant coefficient in Ohio has a slightly larger standard error estimate. Because TX indicator is an example of time-dependent covariate with no heterogeneity before enrollment, we expect there should be no efficiency gain on average for it. Moreover, because of possible correlation between the covariates, it is also possible that some of the coefficient are estimated by PLAC with larger standard error, but the overall efficiency could still be better than the corresponding conditional approach estimates.

	Conditional				PLAC			
	LHR	SE	HR	p	LHR	SE	HR	p
				Ohio				
Young	−0.808	0.181	0.446	0.000	−0.873	0.172	0.418	0.000
Old	0.352	0.138	1.421	0.011	0.364	0.129	1.440	0.005
White	0.328	0.139	1.388	0.019	0.403	0.133	1.497	0.002
Male	0.129	0.128	1.137	0.313	0.141	0.123	1.152	0.249
Treatment								
HD	–	–	–	–	–	–	–	–
PD	0.387	0.217	1.472	0.075	0.466	0.187	1.594	0.013
TX	−1.384	0.144	0.251	0.000	−1.339	0.151	0.262	0.000
				West Virginia				
Young	−1.269	0.592	0.281	0.032	−1.486	0.555	0.226	0.007
Old	1.175	0.473	3.237	0.013	1.140	0.409	3.126	0.005
White	0.703	0.602	2.019	0.243	0.744	0.511	2.103	0.146
Male	0.198	0.401	1.219	0.622	0.204	0.354	1.226	0.564
Treatment								
HD	–	–	–	–	–	–	–	–
PD	1.288	0.955	3.624	0.178	1.580	0.784	4.853	0.044
TX	−1.514	0.489	0.220	0.002	−1.553	0.451	0.212	0.001

Table 4.3: Coefficient estimates from the OPTN/UNOS data in Ohio and West Virginia. PD: peritoneal dialysis; TX: Transplant; and the reference level is hemodialysis (HD). LHR: log hazards ratio; HR: hazards ratio; SE: standard error of LHR; p: p-value (.000: p-value smaller than .001).

4.5 Discussion

The Cox model allows for time-varying covariates and coefficients without major modification of the inferential procedure. The study of Stanford heart transplant program is among the earliest applications of the Cox model with time-dependent covariates (Crowley and Hu, 1977). As a approach to circumvent the left-truncation, the authors would use the enrollment time, i.e., the time a patient is listed, as the time origin for survival analysis (Crowley and Hu, 1977; McDonald and Russ, 2002). In their analysis, the complication due to left-truncation was circumvented by choosing the enrollment as the time of origin. However, because study entry is often irrelevant to the disease progression, the inferential results might be misleading (Thiébaut and Bénichou, 2004). The ESRD onset is a sensible time of origin when analyzing the OPTN/UNOS data, yet we had to deal with the left-truncation and the time-dependent treatment directly.

To this end, we have proposed a semi-parametric estimator for the Cox model with left truncated data involving time-dependent covariates. We have used a pairwise likelihood to eliminated the unspecified truncation distribution, and incorporated the additional information about the parameters of interest to obtain a more efficient estimator. The proposed estimator has appealing large sample properties including a closed-form variance estimator. In numerical studies, it was shown that the efficiency gain of the proposed estimator was larger when the censoring rate is higher and when subjects entered into the sample at a later time.

The greatest challenge in the OPTN/UNOS kidney transplant data was the dependence between the time-varying treatment and the underlying truncation time, which violated the crucial assumption for the pairwise likelihood argument. Nev-

ertheless, this dependence structure possess a certain form such that the proposed estimator can be modified to still yield consistent and more efficient estimates. General dependence structures between $\mathcal{Z}$ and A^* warrant further investigation, where alternative efficiency improvement methods might be necessary.

In some case, studying the effect of time-dependent covariates in the presence of left truncation may face the problem of little or no information of the covariates history before enrollment. When the time-dependent covariates are internal, joint modeling of longitudinal biomarker and time-to-event poses great theoretical and computational challenges (Su and Wang, 2012). Extending our pairwise likelihood augmented estimator to internal time-dependent covariates might be interesting since nowadays prevalent cohort studies often include longitudinal follow-ups.

Our methods can be readily used as a tool to conduct model diagnosis to check the proportional hazards (PH) assumption for the covariates. In the conditional approach, this assumption can be tested by creating an artificial time-dependent covariate (vintage) and test its interaction with the predictor of interest (Kalbfleisch and Prentice, 2002). Similarly, the Wald type test of the same interaction by using the PLAC estimator can be carried out, and it will be more powerful to detect the deviation from the proportionality.

When the PH assumption is violated, time-varying effects can be estimated by using smoothing splines or kernel methods for the conditional approach (Zucker and Karr, 1990; Murphy and Sen, 1991; Tian et al., 2005). Since it is known that the relative risk of PD vs HD changes with the follow-up time, it worth considering a time-varying coefficient model (Vonesh et al., 2006). Our preliminary results using regression splines have shown that the proposed method can give some improvement over the corresponding conditional approach estimator for the time-varying effects.

The corresponding PLAC estimator might also be derived similarly, however, the theoretical justification and the computation burden caused by the non-parametric form of the effects calls for further investigation.

CHAPTER V

Longitudinal Data Clustering Using Penalized Least Squares

5.1 Introduction

Longitudinal data, containing repeated measurements on the same individuals, are valuable for studying either the pathological course of diseases or the normative aging process. Compared with cross-sectional data consist of measurements from different subjects, they typically permit more appreciate inferences on the trend of changes by adjusting for the within-subject correlation. In epidemiological and clinical studies, longitudinal data are often unbalanced, that is, observation times are not common for all subjects. The unbalancedness may be caused by subjects missing visits or dropping out prematurely. Even though the observation times are common by design with no missing or attrition, using a different time scale, e.g., age, will lead to inherent unbalancedness in the analysis. Moreover, the measurements are usually contaminated with random errors, which adds to the difficulty to identify the underlying structures.

Cluster analysis, or clustering, is the art of identifying homogeneous groups from a sample without knowing the labels *a priori* (James and Sugar, 2003). It is one of the most common cognitive activities of human beings to perceive the world. The last decade has seen much development in clustering methods for longitudinal and

functional data, of which Jacques and Preda (2014) provide a taxonomy. Following their terminologies, raw-data methods treats longitudinal data as multivariate. A tremendous amount of literature exists on multivariate clustering methods in pattern recognition, machine learning and statistics (Jain, 2010). Some of the methods can be applied to balanced longitudinal data, although they disregard the intrinsic time ordering of the observations. For unbalanced data, clustering is more challenging. To account for the sparsity and unbalancedness, trajectories are represented by basis functions or principal components assuming their smoothness in time, and the basis expansion coefficients or principal component scores are then assumed to possess subgroup structures. Depending on whether the coefficients are treated as fixed parameters or random variables, these methods are called filtering or adaptive methods. For example, in the filtering step, functional principal component analysis by Yao et al. (2005) can be applied, where the sparsity and unbalancedness are circumvented by local linear smoothers and the conditional expectation of Gaussian variables. Then the principal component scores can be clustered using mixture model for multivariate data (Biernacki et al., 2006). On the other hand, a popular adaptive method for longitudinal data clustering was proposed by James and Sugar (2003) using a reduced rank mixture mixed effect model. They projected the observations on the natural cubic spline basis with identification constrains to overcome the unbalancedness in the sparse observations. An expectation-maximization procedure maximizes the likelihood of all parameters treating the coefficients as unobserved random effects. Lastly, using the essence of cluster analysis, the fourth category of methods rely on properly defined distances. Inspired by Yao et al. (2005), Peng and Müller (2008) defined a distance for unbalanced longitudinal data based on conditional expectations. Common k-means clustering was then used to cluster the

projections from multidimensional scaling.

The Normative Aging Study (NAS) is a longitudinal study of aging in healthy men initiated by the Department of Veterans Affairs in 1963 (Bell et al., 1972). The focus of the study has been on non-pathological aging and the differences between normal and disease-related aging processes (Aldwin et al., 2001). In the NAS, men reported every three to five years for exams including standardized physical and medical exam, clinical chemistry, anthropometric measurements, and medication history. Questionnaires were mailed to assess psychosocial and behavioral topics such as nutrition intake, work and retirement, personality and well-being (Markides, 2007). To study the relationship between normal aging and the natural history of chronic diseases, identifying homogeneous subgroups and the corresponding patterns of the trajectories would be an important exploratory step, for these identified subgroups can in turn serve as benchmarks to classify the subjects, study their association with the risk factors, and aid the decision for intervention and treatment in the era of personalized medicine. When age of the subject is used as the time scale, the repeated measurements in the NAS are highly unbalanced, since the subjects have considerable variability in their ages at entry as well as the between-visit intervals. Therefore, to conduct cluster analysis on the NAS data, methods which can handle unbalanced longitudinal data have to be used.

Convex clustering, a convex relaxation of k-means or hierarchical clustering, was proposed recently (Lindsten et al., 2011; Hocking et al., 2011). Compared with most clustering methods that suffer from local optimum and the need of pre-specified number of clusters, the advantages of convex clustering include continuous clustering path and unique global optimum (Zhu et al., 2014; Chi and Lange, 2015; Tan et al., 2015). It reformulates clustering as an optimization problem with *fusion penalty*

on the norm of pairwise differences of the centroids, i.e., the centers of the clusters. Given $x_1, \ldots, x_m$ in $\mathbb{R}^p$ and their attached centroids $c_1, \ldots, c_m$, the clustering algorithm minimizes the objective function

$$F_\gamma(\mathbf{C}) = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_i\|_2^2 + \gamma \sum_{i<j} w_{ij}\|c_i - c_j\|, \tag{5.1}$$

where the matrix $\mathbf{C}$ has column c_i's, $\|\cdot\|_2$ is the Euclidean norm, γ is a non-negative tuning parameter, w_{ij}'s are non-negative weights. The norm $\|\cdot\|$ for the pairwise differences can be arbitrary (Chi and Lange, 2015), but here we focus on L_1-penalty for simplicity. The convex clustering has been extended to allow for simultaneous feature clustering or selection by introducing additional penalties (Chi et al., 2016; Wang et al., 2016). Similar ideas with concave fusion penalties were also proposed in the subgroup and treatment heterogeneity analysis (Ma and Huang, 2016a,b). Nevertheless, the current study of convex clustering is restricted to multivariate data, where the measurements are assumed less pruned to random errors.

In this project, we propose to extend the clustering method with fusion penalty to unbalanced longitudinal data coarsen by random errors. We approximate the observed trajectories with finite basis functions to address the unbalancedness in the data. The squared differences between the observations and the basis function expansions is minimized, with a fusion penalty on the cluster centers. To define the centers, we evaluate the basis expansions at quantiles of pooled observation times. A related formulation using mixed effect models is introduced, which includes a quadratic penalty on the random effects in addition to the fusion penalty. Alternating direction method of multiplier (ADMM) is used to solve the optimization problem (Boyd et al., 2011). Simulations show the proposed method is robust to various within-cluster heterogeneity and magnitudes of random errors. The proposed method outperforms the existing clustering methods, when clusters mainly differ in shapes,

or when the observations are sparse. Comparison between the two formulations suggests more robust estimation when taking into account of the correlations in the mixed-effect formulation. The application of the proposed method to the NAS study identifies possible subgroups, of which the subjects characteristics are compared.

The rest of this chapter is organized as follows. The proposed clustering method for longitudinal data is presented in Section 5.2. Section 5.3 contains numerical results showing the clustering performance of the proposed methods and those of the existing clustering methods for longitudinal data. The analysis of health trajectories from the NAS cohort is reported in Section 5.4. We conclude the chapter with a brief discussion of the proposed method and future directions in Sections 5.5.

5.2 Proposed Method

5.2.1 Clustering Using Penalized Least Squares

Suppose we have repeated measurements of a continuous outcome from a sample of m subjects. For subject i, denote the outcome at time t_{il} by y_{il}, $i = 1, \ldots, m$, $l = 1, \ldots, n_i$. We approximate the observations by expansions on finite basis functions, $s(t) = (s_1(t), \ldots, s_p(t))^{\mathrm{T}}$, e.g., polynomials or natural cubic splines. The expansion coefficients for subject i is β_i, and we assume the observations

$$y_{il} = \beta_i^{\mathrm{T}} s(t_{il}) + e_{il}; \quad i = 1, \ldots, m,\ l = 1, \ldots, n_i, \tag{5.2}$$

where $\beta_i \in \{b_1, \ldots, b_K\}$, $K \ll m$, and e_{il}'s are i.i.d. with $\mathrm{E}(e_{il}) = 0$ and $\mathrm{Var}(e_{il}) < \infty$. Let Q denote the $p \times q$ matrix of $s(t)$ evaluated at q equally spaced quantiles of the pooled sampled times $(t_{11}, \ldots, t_{1n_1}, \ldots, t_{m1}, \ldots, t_{mn_m})$. The difference between centroids i and j is then $Q^{\mathrm{T}}(\beta_i - \beta_j)$. Given $w_{ij} \geqslant 0$ and $\gamma \geqslant 0$, similar to (5.1), we formulate clustering of longitudinal observations as a least squares problem with a fusion penalty that sums up the L_1 norms of the pairwise differences of centroids

over β_i's:

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}\|y_{il}-\langle\beta_i,s(t_{il})\rangle\|_2^2+\gamma\sum_{i<j}w_{ij}\,\|Q^{\mathrm{T}}(\beta_i-\beta_j)\|$$

where $\langle v_1, v_2\rangle$ is the inner product of vectors v_1 and v_2. Note that the coefficients are not separable because of the penalty term, and hence direct minimization is not easy. Equivalently, we can minimize

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}\|y_{il}-\langle\beta_i,s(t_{il})\rangle\|_2^2+\gamma\sum_{i<j}w_{ij}\,\|\alpha_{ij}\|\,, \tag{5.3}$$

subject to $\alpha_{ij}-Q^{\mathrm{T}}(\beta_i-\beta_j)=0$. By this seemingly redundant constraint, we are able to split the variables in the two terms of (5.3), and update them separately using an alternating direction method of multiplier (ADMM; Boyd et al., 2011). Details of the ADMM algorithm is given in Appendix C.

In effect, (5.2) assumes the observations of the same subject are independent. To appreciate the correlations between observations, assume the expansion coefficients can be decomposed into a fixed effect β_i and a random effect u_i:

$$y_{il}=\beta_i^{\mathrm{T}}s(t_{il})+u_i^{\mathrm{T}}s(t_{il})+e_{il};\quad i=1,\ldots,m,\ l=1,\ldots,n_i, \tag{5.4}$$

where β_i and e_{il}'s satisfy the same conditions as in (5.2), $u_i\sim N(0,G)$, and G is a symmetric positive definite matrix, which we assume is known for other sources. Clustering is then conducted through a penalized least squares with a fusion penalty on the fixed effects and a quadratic penalty on the random effects (cf. Bates and DebRoy, 2004), that is, we minimize

$$\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}\|y_{il}-\langle\beta_i,s(t_{il})\rangle-\langle u_i,s(t_{il})\rangle\|_2^2+\gamma\sum_{i<j}w_{ij}\,\|\alpha_{ij}\|+\frac{1}{2}\sum_{i=1}^{m}u_i^{\mathrm{T}}G^{-1}u_i, \tag{5.5}$$

subject to $\alpha_{ij}-Q^{\mathrm{T}}(\beta_i-\beta_j)=0$. Note that β_i and u_i in (5.5) are not identifiable without the penalties. The fusion penalty makes sure there is not too many distinct centroids, whereas the quadratic penalty shrink them towards the null. Since both

(5.3) and (5.5) take the form of a penalized least squares problem, we name the proposed method clustering using penalized least squares (CUPLS).

5.2.2 Cluster Assignment

In our analysis, we assume each subject belongs to one and only one cluster. We fit (5.2) to get the predicted $\tilde{\beta}_i$ from a linear mixed effect model, and calculate the weights by $w_{ij} = \iota_{ij}^k \exp(-\phi \|\tilde{\beta}_i - \tilde{\beta}_j\|_2^2)$, where ι_{ij}^k is the indicator for $\tilde{\beta}_j$ being within the k nearest neighbors of $\tilde{\beta}_i$ or vice versa, and $\phi > 0$ tunes the bandwidth of the Gaussian kernel (Chi and Lange, 2015). The weights affect the behavior of the clustering paths and hence the results.

Similar to convex clustering, pre-determined number of clusters is not necessary in CUPLS, and continuous clustering path can be obtained by increasing γ over a fine grid. Particularly, we use a geometric sequence, since the agglomeration proceeds more slowly towards the end. For given γ, a breadth-first searching algorithm finds the connected edges from the graph induced by the difference matrix $\boldsymbol{\alpha}$ (Chi and Lange, 2015). An edge is placed between a pair if $|\alpha_{ij}|$ is smaller than a threshold. Connected subjects in the graph are assigned to the same cluster. We repeat the same procedure for the next γ until the sample is fused into one or a small number of clusters, depending on the sparsity of the non-zero weights.

It is sometimes necessary to cut the clustering path to obtain a parsimonious cluster assignment. Jung et al. (2003) proposed the *clustering gain*, to choose cluster assignment for hierarchical clustering of multivariate data. It amounts to finding a balance between maximizing the inter-cluster sum of squares and minimizing the intra-cluster sum of squares. Suppose $\mathcal{C}^\gamma = \{C_1, \ldots, C_{K_\gamma}\}$ is the cluster assignment for γ, with cluster sizes $(m_1, \ldots, m_{K_\gamma})$. As in the weight calculation, we use $\tilde{\beta}_i$ to

calculate the clustering gain:

$$\Delta_\gamma = \sum_{k=1}^{K_\gamma} (m_k - 1) \| Q^{\mathrm{T}} (\tilde{\beta}_0 - \tilde{\beta}_0^{(k)}) \|_2^2, \tag{5.6}$$

where $\tilde{\beta}_0^{(k)} = m_k^{-1} \sum_{i \in C_k} \tilde{\beta}_i$, and $\tilde{\beta}_0 = m^{-1} \sum_{i=1}^{m} \tilde{\beta}_i$. Due to random errors, we choose the most parsimonious assignment with Δ_γ close to the maximum clustering gain. To illustrate, we applied CUPLS with $\gamma = \exp(-4 + g)$, $g = 0, 0.3, 0.6, \ldots$, to simulated data with three equal-size clusters (Figure 5.1; cf. Case 1 of simulation setting I). The clustering gain increases promptly at first, when the numbers of clusters also decreases rapidly. After the number of clusters is less than 10, the clustering gain increment slows down, and it reaches the maximum at $\gamma = 0.9$. We choose the most parsimonious assignment from candidates with $\Delta_\gamma \geqslant 0.95 \Delta_{\max}$, which gives the final clustering assignment with three clusters.

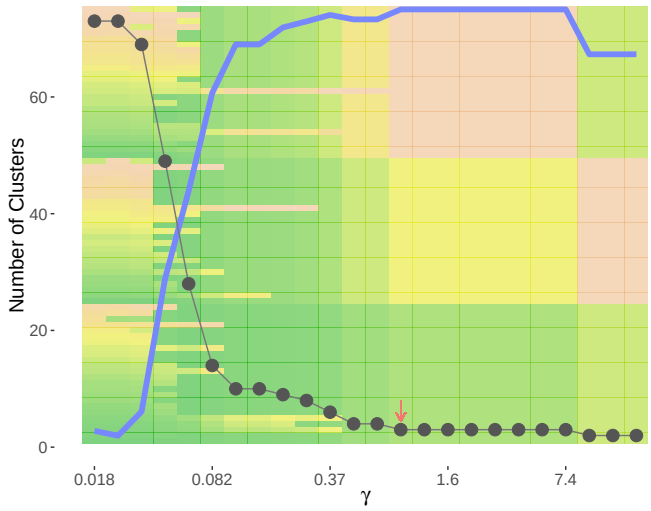

Figure 5.1: An illustration of the use of clustering gain to determine the number of clusters for CUPLS. The colors of the background tiles show the cluster memberships of each subject (row) at each γ (column). The connected dots shows the number of clusters, and the solid blue line indicates the relative magnitude of the clustering gain. The place where the clustering gain reaches the maximum is pointed out by a red arrow.

5.2.3 Comparing Clusterings

We review two criteria measuring the difference between partitions which are invariant to label switching (Stephens, 2000). They will be used in next section to show clustering performance, comparing the clustering results with true memberships.

Suppose there are two partitions of $\{y_1, \ldots, y_m\}$, $\mathcal{C}^1 = \{C_1^1, \ldots, C_{K_1}^1\}$ and $\mathcal{C}^2 = \{C_1^2, \ldots, C_{K_2}^2\}$. Let a and d be the number of pairs clustered together in the same cluster and the number of pairs separated into different clusters in both $\mathcal{C}^1$ and $\mathcal{C}^2$; whereas b and c are the numbers of pairs that are clustered together in one of the partitions, but are separated in the other. The Rand index $R = (a+d)/(a+b+c+d) = (a+d)/\binom{m}{2}$ (Rand, 1971). Hubert and Arabie (1985) adjusted R assuming a hypergeometric distribution for the partitions when they are assigned randomly. The adjusted Rand index (ARI), $(R - \mathrm{E}R)/(1 - \mathrm{E}R)$ lies between -1 and 1, where 1 means the two clusters are identical.

Let $P(k_1) = |\mathcal{C}_{k_1}^1|/m$, $P(k_2) = |\mathcal{C}_{k_2}^2|/m$, and $P(k_1, k_2) = |\mathcal{C}_{k_1}^1 \cap \mathcal{C}_{k_2}^2|/m$, where $|\mathcal{C}|$ is the cardinality of the set $\mathcal{C}$. The variation of information (VI) between $\mathcal{C}^1$ and $\mathcal{C}^2$ is (Meilă, 2007)

$$\begin{aligned} VI(\mathcal{C}^1, \mathcal{C}^2) &= H(\mathcal{C}^1) + H(\mathcal{C}^2) - 2I(\mathcal{C}^1, \mathcal{C}^2) \\ &= -\sum_{k_1=1}^{K_1} P(k_1) \log P(k_1) - \sum_{k_2=1}^{K_2} P(k_2) \log P(k_2) \\ &\quad - 2\sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} P(k_1, k_2) \log \left\{ \frac{P(k_1, k_2)}{P(k_1)P(k_2)} \right\}. \end{aligned}$$

where $H(\mathcal{C}^1)$ and $H(\mathcal{C}^2)$ are the entropies of $P(k_1)$ and $P(k_2)$, respectively, whereas $I(\mathcal{C}^1, \mathcal{C}^2)$ is the mutual information between $\mathcal{C}^1$ and $\mathcal{C}^2$. VI lies between 0 (when the two partitions coincide) and $\log m$ (when one clustering has only one cluster while the other has m), with smaller values indicating better match.

5.3 Simulation

We conducted numerical studies to compare CUPLS with the existing clustering methods for longitudinal data. The methods we considered include the functional clustering model by James and Sugar (2003), the distance-based method by Peng and Müller (2008), and the mixture model to cluster the functional principal component scores (Yao et al., 2005; Biernacki et al., 2006). These methods are denoted as JS, PM and MIX, respectively. The proposed methods, with objective functions (5.3) and (5.5) are denoted as CUPLS and CUPLS_{G} .

We generated longitudinal observations from a linear mixed effect model with quadratic basis functions:

$$y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + e_{ij},\ i = 1, \ldots, m,\ j = 1, \ldots, n_i.$$

The true clustering were $K = 3$ equal-size ($m_k = m/3$) clusters, where for $i \in C_k$, $k \in \{1, \ldots, K\}$, the random coefficients $\beta_i \sim N(b_k, G)$, and $e_i \sim N(0, \sigma_e^2 I_{n_i})$. The baseline, t_{i1}, was drawn from $U(0, \tau)$, $0 < \tau < 1$, and the numbers of follow-ups were from Binomial(n_B, p_{Bi}), where $p_{Bi} = 1 - t_{i1}$. To achieve unbalancedness, the between-visit intervals were generated by $0.1 + \delta$, with $\delta \sim N(0, 0.01)$. Under each scenario, we generated 1000 datasets.

For each dataset, we calculated the ARI and VI, comparing the clustering results with the true clustering assignment. In CUPLS and CUPLS_{G} , natural cubic spline basis were used with knots the same as in JS. To compute the weights, we used $k = 20$ and $\phi = 0.5$. The sequence of γ's started from e^{-5} and increased by multiplying $e^{0.4}$ until it reached e^5 or $\hat{\boldsymbol{\beta}}$ converged, whichever came first. The cluster assignment was the most parsimonious one with $\Delta_\gamma \geqslant 0.95\Delta_{\max}$. We set the penalty parameter $\upsilon = 1$. The profiles of the cluster centers and examples of simulated datasets, as well

as detailed settings for the competing methods are given in Appendix C. Note that JS, PM and MIX were fitted with the knowledge of the correct number of clusters.

In the first set of simulation, we investigate the impact of different signal-noise ratios. Let $G = \sigma_u^2 R$, where R is the correlation matrix of the random effects that resembles the correlation structure in our data application. We varied both σ_u and σ_e between 0.2 and 0.4 to designate different within-cluster heterogeneity and magnitudes of random errors. For the distribution of the random coefficient, considered two cases:

Case 1. $b_1^{\mathrm{T}} = (4, 0, -4)$, $b_2^{\mathrm{T}} = (2, -1, 1)$, and $b_3^{\mathrm{T}} = (4, -8, 4)$; R has exchangeable structure with $\rho = 0.5$, where ρ is the off-diagonal element of R;

Case 2. $b_1^{\mathrm{T}} = (4, 0, -4)$, $b_2^{\mathrm{T}} = (4, -4, 0)$, and $b_3^{\mathrm{T}} = (4, -8, 4)$; the top-left element of R equals one, and the rest are all zeros, i.e., a random intercept model.

The cluster centers in Case 1 differed mainly in level, whereas those in Case 2 differed mainly in shape. To generate the sampled times, we set $\tau = 0.3$, $n_B = 8$ in Case 1 and $\tau = 0.1$, $n_B = 10$ in Case 2. The number of subjects per cluster was $m_k = 25$.

Table 5.1 displays the mean ARI and VI from the first set of simulations. In Case 1, when $\sigma_u = \sigma_e = 0.2$, all methods perform good, with average (ARI, VI) close to $(1, 0)$. When σ_u σ_e increase, the performances of CUPLS and $\mathrm{CUPLS_G}$ are worse than that of JS, but still show better robustness than PM and MIX. Under large σ_u and σ_e, CUPLS usually gives finer partitions nested in the true clusters, because it is inclined to agglomerate the subjects locally, but reluctant to increase the intra-cluster variability by further fusing those small partitions. The deteriorated performance of the convex clustering for multivariate data with low signal-noise ratio was also found previously (Tan et al., 2015). However, in Case 2, CUPLS perform the best under all combinations of σ_u and σ_e. Since in this case trajectories from

different clusters have different shape, but have a lot overlap in the levels during the follow-up, the competing methods tend to cluster incorrectly those with similar levels together. In summary, the proposed method is robust to various magnitudes of intra-cluster heterogeneity and random errors, and it perform better when clusters differ from each other by shapes, but have the same levels.

(σ_u, σ_e)	JS	PM	MIX	CUPLS	CUPLS_{G}
			Case 1: Level difference		
(0.2, 0.2)	(0.960, 0.073)	(0.970, 0.117)	(0.934, 0.135)	(0.995, 0.020)	(0.995, 0.018)
(0.2, 0.4)	(0.950, 0.142)	(0.936, 0.238)	(0.884, 0.291)	(0.948, 0.200)	(0.950, 0.194)
(0.4, 0.2)	(0.947, 0.153)	(0.630, 1.178)	(0.819, 0.407)	(0.899, 0.368)	(0.916, 0.309)
(0.4, 0.4)	(0.831, 0.553)	(0.571, 1.341)	(0.610, 0.901)	(0.713, 0.970)	(0.735, 0.909)
			Case 2: Shape difference		
(0.2, 0.2)	(0.990, 0.040)	(0.882, 0.427)	(0.506, 1.070)	(0.992, 0.033)	(0.992, 0.033)
(0.2, 0.4)	(0.852, 0.517)	(0.795, 0.696)	(0.258, 1.608)	(0.864, 0.487)	(0.864, 0.486)
(0.4, 0.2)	(0.660, 1.009)	(0.439, 1.660)	(0.589, 0.927)	(0.959, 0.163)	(0.963, 0.146)
(0.4, 0.4)	(0.491, 1.492)	(0.411, 1.729)	(0.135, 1.831)	(0.646, 1.176)	(0.669, 1.105)

Table 5.1: Mean clustering index under different within-cluster heterogeneity, measurement errors, and coefficient distributions. σ_u is the standard deviation of the random effects used in generating the data; σ_e is the standard deviation of the measurement errors. JS: James and Sugar (2003); PM: Peng and Müller (2008); MIX: Biernacki et al. (2006); CUPLS and CUPLS_{G}: the proposed methods. The numbers in each cell are the average value of the adjusted Rand index and that of the variation of information.

In the second set of simulations, we study the clustering performance of the methods under various sparsity of observations and sample sizes. We reduced numbers of follow-ups n_B and increased the right boundary of the baseline, τ, to obtain sparse observations. We set $\sigma_u = \sigma_e = 0.2$, and the distribution of the random coefficients is the same as Case 1 above. The number of subjects per cluster considered were $m_k = 25$ and $m_k = 50$.

Average performance index are listed in Table 5.2 for the second set of simulations. When the average number of observations decreases and the overlap of the observation windows become smaller, all methods have undermined performance. CUPLS and CUPLS_{G} are more robust to the increased sparsity, and their perfor-

mance are the best amongst the methods compared when $n_B = 4, \tau = 0.6$. On the other hand, as m_k increases, clustering performance becomes better for all methods, yet the proposed method still gives relatively better results under higher sparsity.

n_B	τ	JS	PM	MIX	CUPLS	CUPLS_G
			$K = 3,\ m_k = 50$			
6	0.4	(0.971, 0.099)	(0.924, 0.308)	(0.974, 0.099)	(0.969, 0.138)	(0.970, 0.137)
4	0.6	(0.661, 0.994)	(0.668, 1.103)	(0.732, 0.705)	(0.847, 0.601)	(0.846, 0.604)
			$K = 3,\ m_k = 25$			
6	0.4	(0.937, 0.167)	(0.864, 0.479)	(0.937, 0.162)	(0.958, 0.138)	(0.958, 0.139)
4	0.6	(0.619, 1.068)	(0.572, 1.276)	(0.565, 0.995)	(0.791, 0.757)	(0.791, 0.760)

Table 5.2: Mean clustering performance index of simulations under various sparsity of the observations. n_B is the maximum number of follow-ups for one subject; τ is the right end point of the starting sampled time. JS: James and Sugar (2003); PM: Peng and Müller (2008); MIX: Biernacki et al. (2006); CUPLS and CUPLS_G : the proposed methods. The numbers in each cell are the average value of the adjusted Rand index and that of the variation of information.

The performance of CUPLS and that of CUPLS_G are almost identical for most scenarios considered in Table 5.1 and 5.2. The discrepancy only becomes obvious when the heterogeneity and error are both large ($\sigma_u = \sigma_e = 0.4$) in Table 5.1. Under these scenarios, CUPLS_G outperforms CUPLS slightly, yielding larger ARI and smaller VI. Separate simulations were conducted using the settings as in Table 5.1 to compare this two alternatives. We computed the deviation of $\hat{\boldsymbol{\beta}}$ from the true centers in the final clusterings which coincided with the true partitions (ARI = 1 and VI = 0). When $\sigma_u = \sigma_e = 0.2$, the mean squared errors of $\hat{\boldsymbol{\beta}}$ (averaged over all coefficients) for CUPLS and CUPLS_G are 0.356 and 0.169 under Case 1, and 0.034 and 0.031 under Case 2. The deviations became larger when $\sigma_u = \sigma_e = 0.4$, but the order was the same. The nuance comes from the different formulations of the model to approximate the trajectories, and the second penalty term on the random effects in CUPLS_G seems to help stabilize $\hat{\beta}_i$'s. Nevertheless, the differences of the estimated centers are not big enough to alter the clustering results.

5.4 Data Application

We apply the proposed method to analyze the NAS data. Since we were interested in the possible non-linear change of the trajectories, we took a subsample of 422 subjects of white race with at least 4 non-missing observations of systolic blood pressure (SBP) and diastolic blood pressure (DBP). The age of the included sample ranged from 23 to 56 years at the study entry, with the median being 37. The cohort were followed for an average of 42 years, and the mean number of follow-ups was 12. Because most observed curves could be approximated by quadratic functions, we applied CUPLS as in (5.3) with quadratic basis expansions. We also clustered the same sample using the model by James and Sugar (2003), assuming the number of clusters was the same as the one chosen by Δ_γ in CUPLS. The observed trajectories and the centers for each cluster are plotted in Figure 5.2 and 5.3 for SBP and DBP, respectively. The cluster centers of the proposed methods were estimated by fitting a linear mixed effect model and the cluster membership as an effect modifier of age.

For SBP, the average trends of the clusters identified by CUPLS and JS are similar. Cluster 1 consists of subjects with the increment accelerated after around 60. Subjects in Cluster 2 started with higher SBP and had an rapid increment of SBP until their sixties, and then the trend reversed afterwards probably due to medication for hypertension. The third cluster is an "average" group, with their average SBP almost the same as the overall mean. However, subjects assigned to the groups by the two methods are different (ARI = 0.33, VI = 2.04). The similarity in trend yet difference in assignment are also observed in the results for DBP. On average, a subject's DBP increases first then decreases after reaching a peak. Two groups are identified; the first one are less healthy, having a higher peak of the increasing phase

and a steeper decrement. The proposed method tends to assign more subjects into this group, yet the agreement between the two methods are closer than that of SBP (ARI = 0.57, VI = 0.88).

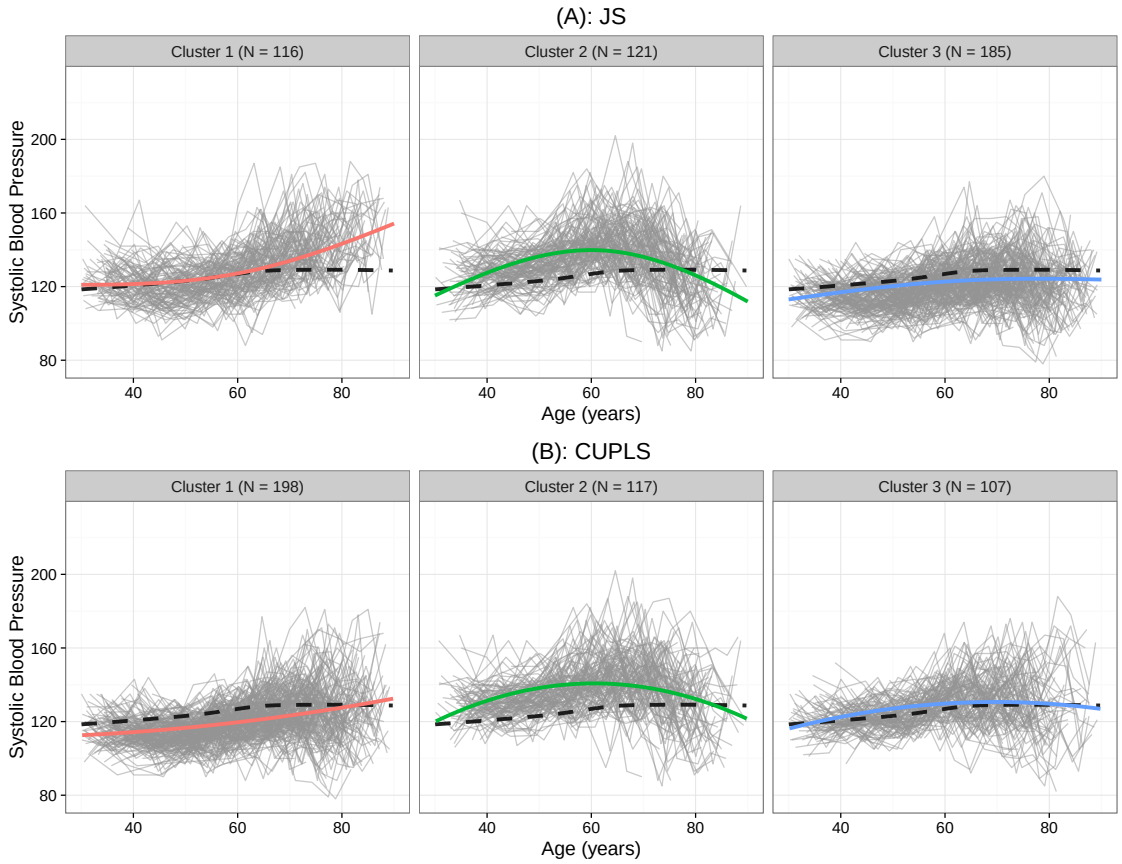

Figure 5.2: Clustering results for SBP using subjects in the NAS subsample with $\geqslant 4$ observations. (A): James and Sugar (2003); (B): CUPLS. The gray lines are the observed trajectories. The solid colored lines is the estimated cluster center (mean trend), whereas the dashed black line is the lowess smoother given the overall mean trend of all trajectories.

The cross table of the cluster assignments using the two methods are given in Table 5.3. With both methods, most subjects in DBP Cluster 1 are also in SBP Cluster 2, who were probably under hypertension medication. The discrepancy between the methods is that CUPLS takes a moderately accelerated increasing trend of SBP (Cluster 1) as "normal", because most of them are in DBP Cluster 2, whereas JS takes the attenuated increasing SBP (Cluster 3) as the normal.

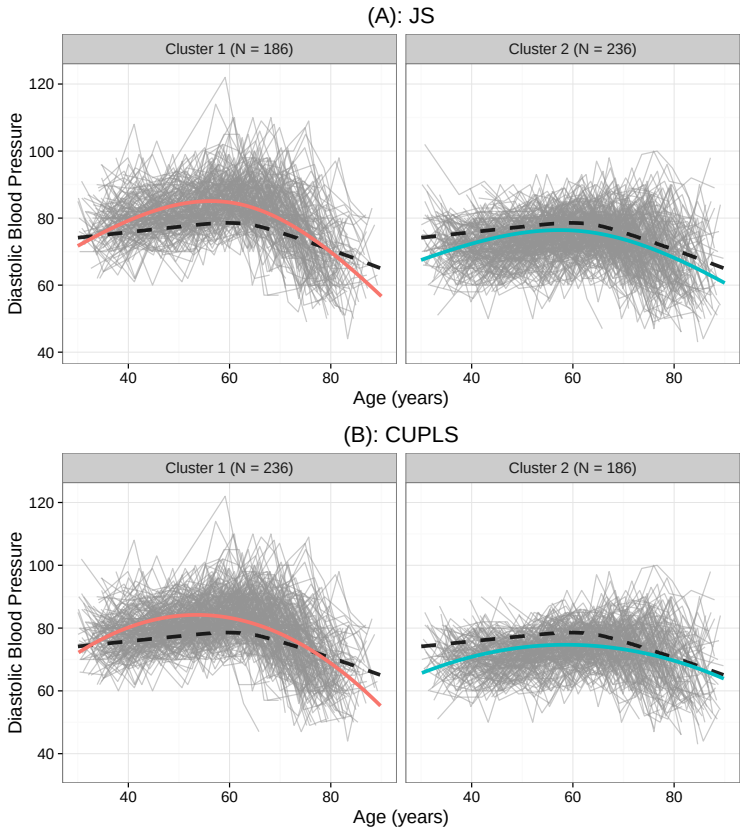

Figure 5.3: Clustering results for DBP using subjects in the NAS subsample with $\geqslant 4$ observations. (A): James and Sugar (2003); (B): CUPLS. The grey lines are the observed trajectories. The solid colored lines is the estimated cluster center (mean trend), whereas the dashed black line is the lowess smoother given the overall mean trend of all trajectories.

		DBP			
		JS		CUPLS	
		1	2	1	2
SBP	1	47	69	47	151
	2	97	24	107	10
	3	42	143	82	25

Table 5.3: Cross table of cluster membership for systolic and diastolic blood pressures. The numbers are the counts. JS: James and Sugar (2003); CUPLS: the proposed method.

We compared the clusters identified by the two methods on demographics (age at entry and education), smoking history, the diagnosis and the onset age of hypertension, and the use of hypertension medication. The smoking history and education were dichotomized into smoker/non-smoker, and had/no college education. No significant differences exists between the three CUPLS SBP clusters in their demographics and smoking history. On the contrary, the SBP clusters identified by JS are more

separate in age at entry ($p = 0.80$) and college education ($p = 0.035$), but smoking history are similar to that in the CUPLS clusters. The comparison of hypertension confirmed our conjecture when only inspecting the trends of SBP clusters. In the second SBP cluster, most of them were diagnosed with hypertension, and more than half of them are on hypertension medication. Moreover, the onset age of the hypertension of this group is also the youngest. For the DBP clustering, Cluster 1 is associated with younger age at entry ($p < 0.001$) and fewer smokers ($p = 0.137$), it is also associate with earlier and higher percentage diagnosis of hypertension as well as higher proportion of hypertension medication usage.

Cluster	SBP				DBP		
	1	2	3	p	1	2	p
	James and Sugar (2003)						
Age at Entry	36.1	36.9	37.7	0.080	36.1	37.8	0.004
Smoking (%)	72	64	68	0.422	62	73	0.012
College (%)	34	20	25	0.035	27	26	0.811
HT Onset Age	58.9	54.6	62.6	<0.001	54.4	64.0	<0.001
HT Diagnosis (%)	71	99	50	<0.001	93	52	<0.001
HT Medication (%)	25	54	24	<0.001	46	23	<0.001
	CUPLS						
Age at Entry	37.3	36.9	36.7	0.518	35.8	38.6	<0.001
Smoking (%)	70	65	68	0.628	65	72	0.137
College (%)	29	25	22	0.392	26	26	0.987
HT Onset Age	65.2	53.6	57.5	<0.001	54.7	67.2	<0.001
HT Diagnosis (%)	46	98	83	<0.001	88	47	<0.001
HT Medication (%)	18	54	38	<0.001	42	22	<0.001

Table 5.4: Demographics and smoking history, and hypertension (HT) comparison for the systolic and diastolic blood pressure clusterings. p-values from Wilcoxon or Kruskal-Wallis rank test are given along with the summary statistics.

5.5 Discussion

The fusion penalty in the proposed method has already been studied earlier in the regularized regression literature to attain sparsity and smoothness (Tibshirani et al., 2005). The difference of our work is that the fusion penalty is used in CUPLS to control the number of different patterns within the *subjects* rather than that of the *features*. The proposed clustering model is similar to the generalized formulation

in Ma and Huang (2016a), where they used concave fusion penalty to capture the latent heterogeneity of the subjects. However, because the main focus in their methods is adjusted for the subgroup in the regression model in stead of clustering, the performance of the algorithm was not sufficiently studied as in our simulation studies. In addition, our method differ from theirs in the repeated longitudinal structure in our data, where quadratic penalty on the random effects can be used to borrow information across subjects.

We have described an ADMM algorithm to implement the proposed method. The CUPLS performed well with clustering close to the truth in the simulation, and yielded reasonable homogeneous subgroups among the NAS sample. We conjecture that The cluster path which connects the centroids from an increasing sequence of γ is continuous (Chi and Lange, 2015). Therefore, we could speed up the ADMM algorithm using *warm start*, i.e., take the solution from the previous objective function as the initial value for the current one. We implemented it in our numerical studies and compared with the same algorithm without warm start. We found using the warm start usually led to less ADMM iterations needed to converge.

As a first step towards thorough investigation of the fusion-penalty-based clustering method, the current work has limitations and further research is warranted. First, the weights for the fusion penalty can be updated using information of the previous centroids on the clustering path. Updating the weights may increase the computation, but weights with more sparsity will compensate that cost. Although work good in application, the weights we used is ad hoc and rely on model fitting. In the future, principal component scores or more generic distances, e.g., Fréchet distance, may be used (Eiter and Mannila, 1994) to calculate the weights. It is also worth to explore the use of other norms, such as smoothly clipped absolute devia-

tions (SCAD; Fan and Li, 2001) and minimax concave penalty (MCP; Zhang, 2010), which may help circumvent the choice of weights as needed in L_1-norm (Ma and Huang, 2016a). Second, as in all penalized methods, selection of appropriate tuning parameters (k, ϕ, γ, ν, ϵ, etc.) is crucial. For simplicity, in our simulation and data example, we did not address this issues and selected the tuning parameters based on empirical criteria as well as the principle of parsimony. Automatic and data driven procedures will make the proposed method easier to implement by the practitioners. Third, in the formulation (5.5), we assumed the covariance structure is known or has been estimated from other sources before we conduct the clustering. In practice, for example, one can first cluster the trajectories using (5.3). We then use the cluster assignment as a effect modifier in a linear mixed effect model to get the estimated $\hat{G}$ for the second penalty.

The proposed method is readily to be extended to including multiple continuous longitudinal outcomes by stacking the observations from different outcomes and the corresponding expansion coefficients (cf. James and Sugar, 2003). In the same manner, incorporating other covariates which may be time-invariant is also possible. Lastly, since non-continuous longitudinal observations, such as those in medication history or quality of life questionnaires, become more prevalent nowadays, extension of the proposed method to outcomes which are counts or categorical data warrants further research.

CHAPTER VI

Conclusions and Future Work

In this dissertation, we have proposed several new statistical methodologies for the analysis for complex survival data and longitudinal data encountered in observational studies. Our goal is to make an efficient use of all information from the available data under plausible assumptions. In addition, we want the proposed methods to be equipped certain robustness to retain reasonable performance when the scenario varies. We have achieved the efficiency and the robustness through the combination of modern inferential tools such as the composite likelihood and the computation algorithms like alternating direction method of multipliers.

In Chapter III, we proposed a more efficient estimator, PLAC, for the Cox model under weak distributional assumption on the truncation distribution. A composite likelihood was constructed, consisting of the conditional likelihood as used in the conventional approaches and a pairwise likelihood as a surrogate for the marginal information of the parameters in the Cox model. The proposed estimator could achieve substantial estimation efficiency gains compared with the conditional approach, even if it does not assume any specific form of truncation beyond its independence of the covariates. The PLAC estimator was shown to have appealing asymptotic properties, and its closed-form variance estimator facilitate the inferences for the related

quantities other than the coefficients and the baseline hazard function.

In Chapter IV, the PLAC estimator was generalized to incorporate time-dependent covariates, and was applied to the OPTN/UNOS kidney transplantation registry data. Different biased sampling scheme was investigated in the simulation studies to show the robustness of the proposed estimator, and provide guidelines to method choices. The violation of the independence assumption of the motivating data was resolved by the modified pairwise likelihood. Application to the kidney transplantation registry data revealed the difference in the long-term survival of ESRD patients under different treatments.

The independence assumption required by the PLAC estimator is weaker than most distribution assumptions in the existing literature for efficiency improvement with left-truncated data. Although a graphical tool was suggested in Chapter III to check this assumption, a more formal test with the observed data warrants further research. The inversed probability weighting scheme as used in the graphical tool might provide a good start for a thorough study. Moreover, the proportion hazard assumption in the Cox model may also fails to hold in practice. With the extension of the PLAC estimator in Chapter IV, it is natural to consider the Cox model with time-varying coefficients. This extension is important in the applications to the kidney transplantation data, for the treatment effect usually changes with time for chronic diseases like the ESRD (Heaf et al., 2002).

In Chapter V, a clustering method, CUPLS, was proposed for sparse and irregular longitudinal data using the least squares with fusion penalty on the pairwise differences of the observations. An alternating direction method of multiplier was used to solve the associated optimization problem. An additional penalty term was suggested to stabilize the clustering results under the alternative random effect formula-

tion. Simulation showed that the proposed method can achieve better or comparable performance compared with the existing longitudinal data clustering method. Currently, CUPLS often results in less agglomerative clustering results, especially when the noise level is large. Ensemble methods could be used to repeat the clustering on random splits of the original data, which may lead to better clustering performance and capture the underlying pattern more completely (Topchy et al., 2004).

APPENDICES

APPENDIX A

Proofs, Additional Simulation and Data Analysis for the First Project

A.1 Proofs of the Asymptotic Properties for the Pairwise Likelihood Augmented Cox Estimator

The asymptotic proofs are given under the following regularity conditions, although weaker ones are possible.

(C1) The true regression coefficients vector β_0 lies in the interior of a compact set $B \subset \mathbb{R}^p$. The true cumulative baseline hazard function $\Lambda_0(t)$ is continuously differentiable and strictly increasing on $[0, \tau]$, and satisfies $\Lambda_0(0) = 0$.

(C2) The covariates vector Z is bounded almost surely. If there exist a deterministic function $b_0(t)$ and a vector $b \in \mathbb{R}^p$, such that $b_0(t) + b^T Z = 0$ with probability one, then $b_0(t) = 0$ and $b = 0$.

(C3) With probability one, there exists a constant $\delta_1 > 0$ such that $\Pr(A^* < T^* \leqslant A^* + C \mid Z^*, A^*, C) > \delta_1$, $\Pr(A + C \geqslant \tau \mid Z) > \delta_1$, and that $\Pr(T \geqslant \tau \mid Z) > \delta_1$.

(C4) Let $b \in \mathbb{R}^p$, and h be a function with bounded total variation on $[0, \tau]$, then the information operator corresponding to the conditional likelihood

evaluated at $(\beta_0,\, \Lambda_0)$,

$$J_0^C(b,\, h) = \left(\lim_{n\to\infty} \partial U^C(\beta, \Lambda)/\partial(\beta, \Lambda)\Big|_{\beta=\beta_0,\, \Lambda=\Lambda_0} \right)(b,\, h),$$

is invertible.

These conditions are standard assumptions for the Cox model under left-truncation, which are necessary to prove the identifiability of the parameters as well as the existence and uniqueness of the PLAC estimator. The continuity of $\Lambda_0(t)$ facilitates the uniform convergence proof of $\hat{\Lambda}(t)$, and the strictly monotonicity suggests that events can happen at any time during the follow-up. The boundedness assumption in (C2) is important for the uniform convergence proofs for the function classes involved, and the second assumption ensures the covariates are not degenerate and that the parameters are identifiable. The first and second assumptions of (C3) imply that for any covariate pattern, subjects with the events happen between 0 and τ have a positive chance to be observed, i.e., not all of them are censored or truncated; whereas the third assumption implies that some subjects could be still at risk by the end of the study. Putting altogether, (C3) ensures the denominator of $\mathcal{L}_C$ is bounded away from zero. Condition (C4), which is used to show the root of the composite score equations is unique, is adapted from the classic weak convergence proof for the Cox model (see Andersen et al., 1993, Condition VII2.1(e)). If with probability one, there exists a constant $\delta_2 > 0$ such that $\Pr(A^* \geqslant T^* \mid Z^*) > \delta_2$, then $\mathcal{L}_n^P$ is non-degenerate, so that we can attain efficiency gain beyond the conditional approach is necessary. If this additional condition does not hold, $\mathcal{L}_n^P$ will be zero, and the PLAC estimator reduced to the common Cox estimator for right-censored data.

We use Ω to denote the set of all possible observations. For convenience, we

adopt the de Finetti's linear functional notations (Pollard, 2002), where $\mathbb{P}_n$ denotes the empirical measure of the observations $\mathcal{O}_i$, $i = 1, \ldots, n$, P_0 denotes the true probability measure on Ω, and $\mathbb{U}_{n,2}$ is the empirical measure of pairs $(\mathcal{O}_i, \mathcal{O}_j)$ such that $1 \leqslant i < j \leqslant n$.

A.1.1 Identifiability

Lemma A.1. *Under Conditions (C1)-(C3), both $\boldsymbol{\beta}_0$ and Λ_0 are identifiable. Specifically, if there exist parameters $(\boldsymbol{\beta}, \Lambda)$ such that Λ is absolutely continuous with respect to Λ_0, $\ell^C(\boldsymbol{\beta}, \Lambda) = \ell^C(\boldsymbol{\beta}_0, \Lambda_0)$ and that $\ell^P(\boldsymbol{\beta}, \Lambda) = \ell^P(\boldsymbol{\beta}_0, \Lambda_0)$ with probability one under P_0, then we have $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\Lambda = \Lambda_0$, where ℓ^C and ℓ^P are the conditional and pairwise log-likelihood functions, respectively.*

Proof. Denote the density and distribution functions of $\mathbf{Z}$ as $f_{\mathbf{Z}}$ and $F_{\mathbf{Z}}$, respectively. First, suppose we have $\ell^C(\boldsymbol{\beta}, \Lambda) = \ell^C(\boldsymbol{\beta}_0, \Lambda_0)$, i.e.,

$$\begin{aligned} & \int_0^\tau \left(\log \lambda(s) + \mathbf{Z}^T\boldsymbol{\beta}\right) dN(s) - \int_0^\tau Y(s)\lambda(s)e^{\mathbf{Z}^T\boldsymbol{\beta}}ds \\ = & \int_0^\tau \left(\log \lambda_0(s) + \mathbf{Z}^T\boldsymbol{\beta}_0\right) dN(s) - \int_0^\tau Y(s)\lambda_0(s)e^{\mathbf{Z}^T\boldsymbol{\beta}_0}ds \end{aligned}$$

holds almost everywhere under P_0, where $N(t) = \Delta I(t \geqslant T)$ is the counting process for the observed event. By Conditions (C1), (C3), and the fact that the support of A^* includes zero, outside a set with zero probability, for any $0 \leqslant a < u \leqslant \tau$ and any $\mathbf{z}$ in the bounded support of $f_{\mathbf{Z}}$, the equality holds for the case with $N(u-) = 0$ and $N(u) = 1$. Taking anti-log transformation on both sides and rearranging the equation, we have

$$\frac{\lambda(u)}{\lambda_0(u)} = e^{\mathbf{z}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta})} \frac{\exp\left\{(\Lambda(u) - \Lambda(a))e^{\mathbf{z}^T\boldsymbol{\beta}}\right\}}{\exp\left\{(\Lambda_0(u) - \Lambda(a))e^{\mathbf{z}^T\boldsymbol{\beta}_0}\right\}}.$$

Let $a \to 0$, we get

$$\frac{\lambda(u)}{\lambda_0(u)} = e^{\mathbf{z}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta})} \frac{\exp\{\Lambda(u)e^{\mathbf{z}^T\boldsymbol{\beta}}\}}{\exp\{\Lambda_0(u)e^{\mathbf{z}^T\boldsymbol{\beta}_0}\}}. \tag{A.1}$$

Moreover, there exists a sequence $\{u_k\}_{k\geqslant 1}$ converging to 0 from above such that (A.1) holds for almost every $\mathbf{z} \in D_k = \{\mathbf{z} : \Pr(C \geqslant u_k | \mathbf{Z} = \mathbf{z}) > 0\}$. Note that $D_k \uparrow D = \{\mathbf{z} : \Pr(C \geqslant 0 \,|\, \mathbf{Z} = \mathbf{z}) > 0\}$, with $\Pr(D) = 1$ under $F_{\mathbf{Z}}$. As $k \to \infty$, the limit for the right-hand side evaluated at u_k, by the continuity of Λ_0 and absolute continuity of Λ with respect to Λ_0, is $e^{\mathbf{z}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta})}$. The left-hand side must also converge, but the limit is independent of $\mathbf{z}$. Hence, the variable $\mathbf{z}^T(\boldsymbol{\beta}_0 - \boldsymbol{\beta})$ is degenerate, which by (C2) implies $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

Substituting $\boldsymbol{\beta}$ with $\boldsymbol{\beta}_0$ in (A.1), on a non-empty set of $\mathbf{z}$ such that $A + C \geqslant \tau$, we have the equality

$$-\lambda(u)e^{\mathbf{z}^T\boldsymbol{\beta}_0} \exp\{-\Lambda(u)e^{\mathbf{z}^T\boldsymbol{\beta}_0}\} = -\lambda_0(u)e^{\mathbf{z}^T\boldsymbol{\beta}_0} \exp\{-\Lambda_0(u)e^{\mathbf{z}^T\boldsymbol{\beta}_0}\}$$

holds for almost every $u \in [0, \tau]$. Integrating both sides from 0 to t yields

$$\int_0^t d\exp\{-\Lambda(u)e^{\mathbf{z}^T\boldsymbol{\beta}_0}\} = \int_0^t d\exp\{-\Lambda_0(u)e^{\mathbf{z}^T\boldsymbol{\beta}_0}\}$$

for all $t \leqslant \tau$. Therefore, we have $\Lambda(t) = \Lambda_0(t)$ for all $t \in [0, \tau]$.

For the pairwise likelihood, outside a set with zero probability, by Condition (C1), for almost every pair of $(\mathbf{z}_1, \mathbf{z}_2)$ in the support of f_Z and every pair of $(A_1,\ A_2)$ such that $\Lambda_0(A_1) - \Lambda_0(A_2) > 0$, we have

$$\frac{\Lambda(A_1) - \Lambda(A_2)}{\Lambda_0(A_1) - \Lambda_0(A_2)} = \frac{e^{\mathbf{z}_1^T\boldsymbol{\beta}_0} - e^{\mathbf{z}_2^T\boldsymbol{\beta}_0}}{e^{\mathbf{z}_1^T\boldsymbol{\beta}} - e^{\mathbf{z}_2^T\boldsymbol{\beta}}}. \tag{A.2}$$

Thus (A.2) implies that the ratios on both sides are the same constant c. By Condition (C1), the left-hand side then gives $\Lambda(t) = c\Lambda_0(t)$ for t in the support of A. On the other hand, the right-hand side is degenerate if it equals c when $(\mathbf{z}_1, \mathbf{z}_2)$ vary, this again implies $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ thus $c = 1$. Note that we need the support of A includes the follow-up period to identify Λ_0 solely from the pairwise likelihood.

It is worth noting that $\Lambda_0(t)$ is not identifiable for $0 < t < w_1$ (Wang et al., 1993). However, since the support of A^* includes zero, by (C3), w_1 is usually close to zero; thus, the identifiability issue is less likely to occur. □

A.1.2 Consistency

The PLAC estimator falls in the category of Z-estimators. To follow the consistency proof of the general Z-estimators, a complication brought by the pairwise structure is to show the uniform convergence of the involved bivariate function classes. We tackle this difficulty through bounding the bracketing numbers (entropies) of these function classes using the U-processes theory (see De la Peña and Giné, 1999, Chapter 5). For $k = 0, 1, 2$, the function classes $\{(\mathbf{z}_1, \mathbf{z}_2) \mapsto \mathbf{z}_1^{\otimes k} e^{\mathbf{z}_1^T \boldsymbol{\beta}} - \mathbf{z}_2^{\otimes k} e^{\mathbf{z}_2^T \boldsymbol{\beta}} : \mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^p; \boldsymbol{\beta} \in B\}$ are Euclidean (Nolan and Pollard, 1987); thus, their bracketing numbers in $L_1(P^2)$ are finite, where $P^2 \equiv P \otimes P$, and P is any probability measure. Bounds for classes only consisting of indicator functions can be shown using the VC theory (see De la Peña and Giné, 1999, Section 5.2). Denoting the class of cumulative baseline hazard functions satisfying (C1) as $\mathcal{H}_\Lambda$, then

Lemma A.2. *The function class* $\mathcal{H}_\Lambda^D = \{(s,t) \mapsto \Lambda(s) - \Lambda(t) : s, t \in [0, \tau]; \Lambda \in \mathcal{H}_\Lambda\}$ *has finite bracketing numbers in* $L_1(P^2)$ *for all* $\varepsilon > 0$.

Proof. To avoid technicality, we assume all bivariate function classes involved in this and the following proofs are measurable (see De la Peña and Giné, 1999, Section 3.5). Theorem 2.7.5 of van der Vaart and Wellner (1996) indicates that for a fixed $\varepsilon > 0$, there exists a constant K_1 such that the bracketing entropy

$$\log N_{[]}(\varepsilon, \mathcal{H}_\Lambda, L_1(P)) < \frac{K_1}{\varepsilon} < \infty$$

for any probability measure P. For a given $\Lambda \in \mathcal{H}_\Lambda$, suppose an ε-bracket containing

it in $L_1(P)$ is (Λ_l, Λ_u); thus, we have $\Lambda_l(t) < \Lambda(t) < \Lambda_u(t)$, $\forall t \in [0, \tau]$ and that

$$\int |\Lambda_u(s) - \Lambda_l(s)| \, dP < \varepsilon.$$

Then for the corresponding bivariate function in $\mathcal{H}_\Lambda^D$, we have

$$\Lambda_l(s) - \Lambda_u(t) < \Lambda(s) - \Lambda(t) < \Lambda_u(s) - \Lambda_l(t), \quad \forall s, t \in [0, \tau] \, .$$

By triangle inequality,

$$\begin{aligned} & \iint |\Lambda_u(s) - \Lambda_l(t) - \Lambda_l(s) + \Lambda_u(t)| \, dP^2 \\ \leq & \int \int |\Lambda_u(s) - \Lambda_l(s)| \, dPdP + \int \int |\Lambda_u(t) - \Lambda_l(s)| \, dPdP \\ = & \int |\Lambda_u(s) - \Lambda_l(s)| \, dP + \int |\Lambda_u(t) - \Lambda_l(s)| \, dP < 2\varepsilon. \end{aligned}$$

Therefore, $(\Lambda_l(s) - \Lambda_u(t), \, \Lambda_u(s) - \Lambda_l(t))$ is a 2ε-bracket for $\Lambda(s) - \Lambda(t)$ in $L_1(P^2)$, thus there is a constant $K_2 > 0$ such that the bracketing entropy

$$\log N_{[]}(\varepsilon, \mathcal{H}_\Lambda^D, L_1(P^2)) < \frac{K_2}{\varepsilon} < \infty.$$

Since ε is arbitrary, the class $\mathcal{H}_\Lambda^D$ has finite bracketing numbers in $L_1(P^2)$. □

Remark A.3. By Corollary 5.2.5 of De la Peña and Giné (1999), the finite bracketing numbers imply the corresponding function classes satisfy the uniform law of large numbers of U-processes. The uniform law of large numbers for $U^P(\boldsymbol{\beta}, \Lambda)$ and its derivatives then follows, because they are Lipschitz functions of the component functions with finite bracketing numbers (van der Vaart and Wellner, 1996).

Proof of Theorem III.1. We first re-write the modified composite log-likelihood (3.2) and the composite score functions using the linear functional notations. Let $N_i(s) = \Delta_i I(X_i \leqslant s)$ be the observed event counting process for subject i, then (3.2) can be

written as

$$\begin{aligned}\ell_n^c(\boldsymbol{\beta},\Lambda) &= \mathbb{P}_n\int_0^\tau \left\{(\log\Lambda\{s\}+\mathbf{Z}^T\boldsymbol{\beta})dN(s)-Y(s)e^{\mathbf{Z}^T\boldsymbol{\beta}}d\Lambda(s)\right\}\\ &\quad -\mathbb{U}_{n,2}\log(1+R(\boldsymbol{\beta},\Lambda)).\end{aligned}$$

Differentiating it with respect to $\boldsymbol{\beta}$ yields the composite score function for $\boldsymbol{\beta}$:

$$\begin{aligned}U_{\boldsymbol{\beta}}(\boldsymbol{\beta},\Lambda) &= \mathbb{P}_n\int_0^\tau \mathbf{Z}\left\{dN(s)-Y(s)e^{\mathbf{Z}^T\boldsymbol{\beta}}d\Lambda(s)\right\}\\ &\quad -\mathbb{U}_{n,2}\left\{\frac{R(\boldsymbol{\beta},\Lambda)}{1+R(\boldsymbol{\beta},\Lambda)}\int_0^\tau Q^{(1)}(s;\boldsymbol{\beta})d\Lambda(s)\right\}.\end{aligned}$$

For $0\leqslant t\leqslant\tau$ and $h(\cdot)=I(\cdot\leqslant t)$, define a perturbation of Λ by $d\Lambda_\varepsilon=(1+\varepsilon h)d\Lambda$. The derivative of $\ell_n^c(\boldsymbol{\beta},\Lambda_\varepsilon)$ with respect to ε evaluated at $\varepsilon=0$ yields the composite score function for Λ in the direction of h:

$$\begin{aligned}U_\Lambda(\boldsymbol{\beta},\Lambda)(h) &= \mathbb{P}_n\int_0^\tau h(s)\left\{dN(s)-Y(s)e^{\mathbf{Z}^T\boldsymbol{\beta}}d\Lambda(s)\right\}\\ &\quad -\mathbb{U}_{n,2}\left\{\frac{R(\boldsymbol{\beta},\Lambda)}{1+R(\boldsymbol{\beta},\Lambda)}\int_0^\tau Q^{(0)}(s;\boldsymbol{\beta})h(s)d\Lambda(s)\right\}.\end{aligned}$$

As in Section 2.3, we can write the composite score function

$$U(\boldsymbol{\beta},\Lambda)=\begin{pmatrix}U_{\boldsymbol{\beta}}(\boldsymbol{\beta},\Lambda)\\ U_\Lambda(\boldsymbol{\beta},\Lambda)(h)\end{pmatrix}$$

as the summation of $U^C(\boldsymbol{\beta},\Lambda)$ and $U^P(\boldsymbol{\beta},\Lambda)$; the former is the conditional approach score function and has expectation zero. We can also show that $\mathrm{E}_0\{U^P(\boldsymbol{\beta}_0,\Lambda_0)\}=0$, since the summand of U^P satisfies $\mathrm{E}_0\{U_{ij}^P(\boldsymbol{\beta}_0,\Lambda_0)\}=0$, $1\leqslant i<j\leqslant n$. To see this, note that the pair $(A_i,\,A_j)$ has a binary distribution after conditioning on $(\mathbf{Z}_i,\mathbf{Z}_j)$ and the order statistics of $(A_i,\,A_j)$; thus, by double expectation, we have

$$\begin{aligned}\mathrm{E}_0\{U_{ij}^P(\boldsymbol{\beta},\Lambda)\} &= \mathrm{E}_0\left\{\frac{1}{1+R_{ij}^{-1}(\boldsymbol{\beta},\Lambda)}\cdot\frac{1}{1+R_{ij}(\boldsymbol{\beta}_0,\Lambda_0)}\begin{pmatrix}\int Q^{(1)}(s;\boldsymbol{\beta})d\Lambda(s)\\ \int h(s)Q^{(0)}(s;\boldsymbol{\beta})d\Lambda(s)\end{pmatrix}\right.\\ &\quad \left.-\frac{1}{1+R_{ij}(\boldsymbol{\beta},\Lambda)}\cdot\frac{1}{1+R_{ij}^{-1}(\boldsymbol{\beta}_0,\Lambda_0)}\begin{pmatrix}\int Q^{(1)}(s;\boldsymbol{\beta})d\Lambda(s)\\ \int h(s)Q^{(0)}(s;\boldsymbol{\beta})d\Lambda(s)\end{pmatrix}\right\}. \qquad \text{(A.3)}\end{aligned}$$

The two terms in the bracket cancel if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\Lambda = \Lambda_0$ by identifiability.

Since $\log \mathcal{L}_n^P$ is always negative, by the similar arguments as in Zeng and Lin (2006), we can show that the PLAC estimator has finite jump sizes, and that $\hat{\Lambda}(\tau)$ is bounded a.s. when $n \to \infty$. Because $\ell_n^c(\boldsymbol{\beta}, \Lambda)$ is maximized at the PLAC estimator $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$ over the whole model, it is certainly maximized along the parametric submodel $(\hat{\boldsymbol{\beta}}, \Lambda_\varepsilon)$ at $\varepsilon = 0$. Thus by the regularity conditions, the PLAC estimator is the solution to the composite score equations $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \Lambda) = 0$ and $U_\Lambda(\boldsymbol{\beta}, \Lambda)(h) = 0$. Interchanging the summations and integrals in the second equation and rearranging the resulting terms, we have

(A.4)
$$\mathbb{P}_n \int_0^\tau h(s) dN(s) = \int_0^\tau h(s) \left\{ \mathbb{P}_n Y(s) e^{\mathbf{Z}^T \hat{\boldsymbol{\beta}}} + \mathbb{U}_{n,2} \frac{R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})}{1 + R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})} Q^{(0)}(s; \hat{\boldsymbol{\beta}}) \right\} d\hat{\Lambda}(s).$$

Let

$$M_n(s; \hat{\boldsymbol{\beta}}, \hat{\Lambda}) = \mathbb{P}_n Y(s) e^{\mathbf{Z}^T \hat{\boldsymbol{\beta}}} + \mathbb{U}_{n,2} \frac{R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})}{1 + R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})} Q^{(0)}(s; \hat{\boldsymbol{\beta}})$$

denote the random function in the brackets. Replacing $h(s)$ with $h(s)/M_n(s; \hat{\boldsymbol{\beta}}, \hat{\Lambda})$ on both sides of (B.3) yields the self-consistency solution of Λ:

$$\hat{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \hat{\boldsymbol{\beta}}, \hat{\Lambda})}.$$

The rest of the proof follows closely to Murphy et al. (1997), yet the technical details are different due to the pairwise pseudo-likelihood. Inspired by the form of $\hat{\Lambda}$, we define another random step function

$$\tilde{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \boldsymbol{\beta}_0, \Lambda_0)}.$$

Let $M_0(s; \boldsymbol{\beta}_0, \Lambda_0) = P_0\{Y(s) e^{\mathbf{Z}^T \boldsymbol{\beta}_0}\}$. Since $\mathrm{E}_0\{U(\boldsymbol{\beta}_0, \Lambda_0)\} = 0$ and $\mathrm{E}_0\{U^P(\boldsymbol{\beta}_0, \Lambda_0)\} =$

0, the same algebra as we used to get $\hat{\Lambda}$ yields

$$\Lambda_0(t) = P_0 \int_0^t \frac{dN(s)}{M_0(s; \boldsymbol{\beta}_0, \Lambda_0)}.$$

Under the regularity conditions (C2)-(C3), by Lemma A.2, and the double expectation argument as we used in (A.3), $s \mapsto M_n(s; \boldsymbol{\beta}_0, \Lambda_0)$ is uniformly bounded away from zero and infinity, and is of uniformly bounded variation when n is sufficiently large. Therefore, by the Glivenko–Cantelli theorem and Remark A.3, we have

$$\|M_n(s; \boldsymbol{\beta}_0, \Lambda_0) - M_0(s; \boldsymbol{\beta}_0, \Lambda_0)\|_{L_\infty[0,\tau]} \overset{a.s.}{\to} 0$$

and

$$\left\| \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s; \boldsymbol{\beta}_0, \Lambda_0)} - P_0 \int_0^t \frac{dN(s)}{M_n(s; \boldsymbol{\beta}_0, \Lambda_0)} \right\|_{L_\infty[0,\tau]} \overset{a.s.}{\to} 0,$$

where $\|\cdot\|_{L_\infty[0,\tau]}$ is the supreme norm over $[0, \tau]$. These results combined with the dominated convergence theorem yield

$$\left\| \tilde{\Lambda}(t) - \Lambda_0(t) \right\|_{L_\infty[0,\tau]} \overset{a.s.}{\to} 0.$$

By the definition of the PLAC estimator, the log-composite-likelihood evaluated at $(\hat{\beta}, \hat{\Lambda})$ is greater than that evaluated at $(\beta_0, \tilde{\Lambda})$:

$$\begin{aligned} &\mathbb{P}_n \int_0^\tau \left\{ \log \frac{\hat{\Lambda}}{\tilde{\Lambda}}\{s\} + \mathbf{Z}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} dN(s) \\ -\mathbb{P}_n &\left\{ e^{\mathbf{Z}^T\hat{\boldsymbol{\beta}}} \int_0^\tau Y(s) d\hat{\Lambda}(s) - e^{\mathbf{Z}^T\boldsymbol{\beta}_0} \int_0^\tau Y(s) d\tilde{\Lambda}(s) \right\} - \mathbb{U}_{n,2} \log \frac{1 + R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})}{1 + R(\boldsymbol{\beta}_0, \tilde{\Lambda})} \quad \geqslant \quad 0. \end{aligned}$$

By assumption, $\boldsymbol{\beta}$ is in a compact set, and that $\hat{\Lambda}(t) \leqslant \hat{\Lambda}(\tau)$ is bounded for $t \in [0, \tau]$ with probability one. Thus, by the Bolzano–Weierstrass theorem and the Helly's selection lemma, for every subsequence of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$, we can find a further subsequence (still denoted as $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$) along which $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^*$ and $\hat{\Lambda}(t) \to \Lambda^*(t)$, $\forall t \in [0, \tau]$ for some monotone function Λ^* almost surely.

Since $\hat{\Lambda}(t)$ is absolutely continuous with respect to $\tilde{\Lambda}(t)$, let $\eta(t) = \lim_{n\to\infty} d\hat{\Lambda}/d\tilde{\Lambda}$ be a bounded measurable function, then $\Lambda^*(t) = \int_0^t \eta(s) d\Lambda_0(s)$ (Zeng and Lin, 2006). By (C1), $\Lambda^*(t)$ is absolutely continuous with respect to the Lebesgue measure and we denote its derivative as $\lambda^*(t)$. Thus we have the ratio $d\hat{\Lambda}/d\tilde{\Lambda}$ converges to $\eta(t) = \lambda^*(t)/\lambda_0(t)$. Again, by the Glivenko–Cantelli theorem, Lemma A.2, Remark A.3 and the dominant convergence theorem, the difference of the log-composite-likelihoods converges to

$$\begin{aligned} & P_0 \int_0^\tau \left\{ \log \frac{\lambda^*}{\lambda_0}(s) + \mathbf{Z}^T(\boldsymbol{\beta}^* - \boldsymbol{\beta}_0) \right\} dN(s) \\ -P_0 \left\{ e^{\mathbf{Z}^T\boldsymbol{\beta}^*} \int_0^\tau Y(s) d\Lambda^*(s) - e^{\mathbf{Z}^T\boldsymbol{\beta}_0} \int_0^\tau Y(s) d\Lambda_0(s) \right\} - P_0 \log \frac{1 + R(\boldsymbol{\beta}^*, \Lambda^*)}{1 + R(\boldsymbol{\beta}_0, \Lambda_0)} & \geqslant 0. \end{aligned}$$

The left-hand side is the composite Kullback–Leibler divergence (Varin and Vidoni, 2005) of the density indexed by $(\boldsymbol{\beta}^*, \Lambda^*)$ from the true density, which by Kullback–Leibler inequality and Lemma A.1 should be strictly negative unless $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\Lambda^* = \Lambda_0$. Since every subsequence of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$ has a further subsequence converging to $(\boldsymbol{\beta}_0, \Lambda_0)$, we have the convergence of the entire sequence to the same limit. Finally, the uniform convergence of $\hat{\Lambda}(t)$ to $\Lambda_0(t)$ over $[0, \tau]$ follows from the continuity of Λ_0. □

A.1.3 Asymptotic Normality

We first establish a lemma on the $\sqrt{n}$-uniform convergence rate and asymptotic normality of the log-generalized odds ratio. This is achieved by the projection of the U-process.

Lemma A.4. *Under Conditions (C1)-(C4), the class of the log-generalized odds ratios*

$$\mathcal{R} = \left\{ (\mathcal{O}_i, \mathcal{O}_j) \mapsto r_{ij}(\boldsymbol{\beta}, \Lambda) : \mathcal{O}_i, \mathcal{O}_j \in \Omega, \boldsymbol{\beta} \in B, \Lambda \in \mathcal{H}_\Lambda \right\},$$

where $r_{ij}(\boldsymbol{\beta}, \Lambda) = (e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - e^{\mathbf{Z}_j^T\boldsymbol{\beta}})(\Lambda(A_i) - \Lambda(A_j))$, *satisfies the uniform central limit theorem for U-processes:*

$$\sqrt{n}(\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda)) \rightsquigarrow \mathbb{G}_r,$$

where $\mathbb{G}_r$ *is a tight mean-zero Gaussian process.*

Proof. To establish the weak convergence, we first show that

$$\left\| \mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda) - \hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda) \right\|_{\boldsymbol{\beta},\Lambda} = o_p(n^{-1/2}),$$

where

$$\hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda) = \sum_{i=1}^{n} \mathrm{E}\left(\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda) | \mathcal{O}_i \right)$$

is the Hájek projection of $\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda)$ (van der Vaart, 2000), and $\|\cdot\|_{\boldsymbol{\beta},\Lambda}$ is the supreme norm over the parameter space.

It can be verified that $P_0^2 r(\boldsymbol{\beta}, \Lambda) = 2\mathrm{Cov}(e^{\mathbf{Z}^T\boldsymbol{\beta}}, \Lambda(A))$. Moreover, since the pair $\mathcal{O}_i$ and $\mathcal{O}_j$ are i.i.d.,

$$\begin{aligned} \mathrm{E}\left(r_{ij}(\boldsymbol{\beta}, \Lambda) \,|\, \mathcal{O}_i \right) &= \mathrm{E}\left\{ (e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - e^{\mathbf{Z}_j^T\boldsymbol{\beta}})(\Lambda(A_i) - \Lambda(A_j)) \,\middle|\, A_i, \mathbf{Z}_i \right\} \\ &= e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_i) - \Lambda(A_i)\mathrm{E}e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\mathrm{E}\Lambda(A_i) + \mathrm{E}(e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_i)). \end{aligned}$$

Thus we have

$$\begin{aligned} \hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda) &= \sum_{i=1}^{n} \mathrm{E}\left\{ \binom{n}{2}^{-1} \sum_{j<k} r_{jk}(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda) \,\middle|\, \mathcal{O}_i \right\} \\ &= \frac{2}{n} \sum_{i=1}^{n} \left\{ e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_i) - \Lambda(A_i)\mathrm{E}e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\mathrm{E}\Lambda(A_i) + \mathrm{E}(e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_i)) \right\} \\ &\quad - 4\mathrm{Cov}(e^{\mathbf{Z}^T\boldsymbol{\beta}}, \Lambda(A)). \end{aligned}$$

Direct calculation gives

$$\tilde{\mathbb{U}}_{n,2} \equiv \mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P^2 r(\boldsymbol{\beta}, \Lambda) - \hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda) = \frac{1}{\binom{n}{2}} \sum_{i<j} \tilde{\mathbb{U}}_{n,2}^{(i,j)}.$$

The summand of $\tilde{\mathbb{U}}_{n,2}$ is given by

$$\begin{aligned}
\tilde{\mathbb{U}}_{n,2}^{(i,j)} &= e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_i) - e^{\mathbf{Z}_j^T\boldsymbol{\beta}}\Lambda(A_i) - e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_j) + e^{\mathbf{Z}_j^T\boldsymbol{\beta}}\Lambda(A_j) - 2\mathrm{Cov}(e^{\mathbf{Z}^T\boldsymbol{\beta}}, \Lambda(A)) \\
&\quad - \left\{e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_i) - \Lambda(A_i)\mathrm{E}e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\mathrm{E}\Lambda(A_i) + \mathrm{E}(e^{\mathbf{Z}_i^T\boldsymbol{\beta}}\Lambda(A_i))\right\} \\
&\quad - \left\{e^{\mathbf{Z}_j^T\boldsymbol{\beta}}\Lambda(A_j) - \Lambda(A_j)\mathrm{E}e^{\mathbf{Z}_j^T\boldsymbol{\beta}} - e^{\mathbf{Z}_j^T\boldsymbol{\beta}}\mathrm{E}\Lambda(A_j) + \mathrm{E}(e^{\mathbf{Z}_j^T\boldsymbol{\beta}}\Lambda(A_j))\right\} \\
&\quad + 4\mathrm{Cov}(e^{\mathbf{Z}^T\boldsymbol{\beta}}, \Lambda(A)) \\
&= -(e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - \mathrm{E}e^{\mathbf{Z}_i^T\boldsymbol{\beta}})\left(\Lambda(A_j) - \mathrm{E}\Lambda(A_j)\right) - (e^{\mathbf{Z}_j^T\boldsymbol{\beta}} - \mathrm{E}e^{\mathbf{Z}_j^T\boldsymbol{\beta}})\left(\Lambda(A_i) - \mathrm{E}\Lambda(A_i)\right),
\end{aligned}$$

where the second equality holds by the definition of the covariance and the i.i.d. property of the observations. Therefore, we have

$$\begin{aligned}
\tilde{\mathbb{U}}_{n,2} &= -\frac{1}{\binom{n}{2}}\sum_{i=1}^{n}\sum_{j=1}^{n}(e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - \mathrm{E}e^{\mathbf{Z}_i^T\boldsymbol{\beta}})\left(\Lambda(A_j) - \mathrm{E}\Lambda(A_j)\right) \\
&\asymp -2\cdot\frac{1}{n}\sum_{i=1}^{n}(e^{\mathbf{Z}_i^T\boldsymbol{\beta}} - \mathrm{E}e^{\mathbf{Z}_i^T\boldsymbol{\beta}})\cdot\frac{1}{n}\sum_{j=1}^{n}\left(\Lambda(A_j) - \mathrm{E}\Lambda(A_j)\right),
\end{aligned}$$

where $\asymp$ means asymptotically equivalent. Since both summations in the last line are empirical processes of Donsker classes, we have

$$\left\|\tilde{\mathbb{U}}_{n,2}\right\|_{\boldsymbol{\beta},\Lambda} \lesssim \left\|n^{-1/2}\mathbb{G}_n e^{\mathbf{Z}^T\boldsymbol{\beta}}\right\|_{\boldsymbol{\beta}}\cdot\left\|n^{-1/2}\mathbb{G}_n\Lambda\right\|_{\Lambda} = O_p(n^{-1/2})O_p(n^{-1/2}) = o_p(n^{-1/2}),$$

where $\lesssim$ means the inequality holds up to a multiplicative constant and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$.

Therefore, $\mathbb{U}_{n,2}r(\boldsymbol{\beta},\Lambda) - P_0^2 r(\boldsymbol{\beta},\Lambda)$ is equivalent to its projection $\hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta},\Lambda)$ up to a term of $o_p(n^{-1/2})$. The weak convergence of $\hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta},\Lambda)$ can be established using the empirical process theory. Combining these two facts leads to the weak convergence of $\mathbb{U}_{n,2}r(\boldsymbol{\beta},\Lambda)$. □

Proof of Theorem III.2. Let θ denote the parameters $(\boldsymbol{\beta},\Lambda)$. We proceed by checking the four conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996). Note that

$\sqrt{n}U(\theta_0)$ can be decomposed into $\sqrt{n}U^C(\theta_0) + \sqrt{n}U^P(\theta_0)$. Following the martingale theory, the first term converges weakly to a mean-zero Gaussian process $\mathbb{G}_{U^C}$, and the linear functional

$$\sqrt{n}\left\{b_1^T U_{\boldsymbol{\beta}}^C(\theta_0) + U_\Lambda^C(\theta_0)(h)\right\}$$

converges weakly to a mean-zero normal random variable with the variance that can be consistently estimated by $b^T\hat{V}^C b$, where b is defined as in Section 3.2.3. For the second term, by Lemma A.4, the preservation theorem of Lipschitz functions and Theorem 5.3.1 of (De la Peña and Giné, 1999), it also converges weakly to a mean-zero Gaussian process $\mathbb{G}_{U^P}$, and the linear functional

$$\sqrt{n}\left\{b_1^T U_{\boldsymbol{\beta}}^P(\theta_0) + U_\Lambda^P(\theta_0)(h)\right\}$$

converges weakly to a mean-zero normal random variable with the variance that can be consistently estimated by $b^T\hat{V}^P b$. Note also that given $\{(A_i, \mathbf{Z}_i)\}_{i=1}^n$, $U^C(\theta_0)$ is a martingale, whereas $U^P(\theta_0)$ is a function of A_i and $\mathbf{Z}_i$ only, thus by the double expectation

$$\begin{aligned}
\mathrm{E}_0\{U^C(\theta_0)\cdot U^P(\theta_0)\} &= \mathrm{E}_0\left\{\mathrm{E}_0\left(U^C(\theta_0)\middle|\,(A_i, Z_i), i=1,\ldots,n\right)\cdot U^P(\theta_0)\right\} \\
&= \mathrm{E}_0\{0\cdot U^P(\theta_0)\} = 0,
\end{aligned}$$

where $\cdot$ denotes the inner product of the underlying space. This indicates that the $U^C(\theta_0)$ and $U^P(\theta_0)$ are asymptotically independent (van der Vaart and Wellner, 1996, Example 1.4.6) at $\theta_0$that $\sqrt{n}U(\theta_0)$ converges weakly to a mean-zero Gaussian process $\mathbb{G}_U$. In addition, $\sqrt{n}\left\{b_1^T U_{\boldsymbol{\beta}}(\theta_0) + U_\Lambda(\theta_0)(h)\right\}$ converges weakly to a mean-zero normal random variable with asymptotic variance that can be consistently estimated by $b^T(\hat{V}^C + \hat{V}^P)\,b$. Therefore, the two stochastic conditions are satisfied by the consistency of $\hat{\theta}$, Lemma A.4 and Lemma 3.3.5 of van der Vaart and Wellner (1996).

The fourth condition holds since $\hat{\theta}$ is a zero of $U(\theta)$ and that $u(\theta_0) \equiv \mathrm{E}_0 U(\theta_0) = 0$ by the arguments in the consistency proof.

To complete the proof, we only need to verify that the Fréchet-derivative of u at θ_0 exists and is continuous invertible. The Fréchet-differentiability can be check directly. For the continuous invertibility, note that the derivative $J \equiv \partial u(\theta)/\partial\theta|_{\theta=\theta_0}$ can be decomposed into J^C and J^P. By (C5) and the classic Cox model results, the first part is continuously invertible. Thus, it suffices to show J^P is a compact operator and that J is one-to-one by the Fredholm theory.

Following Example 3.3.10 of van der Vaart and Wellner (1996), we find the derivate J^P has the form

$$\begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}_0 \\ \Lambda - \Lambda_0 \end{pmatrix} \mapsto \begin{pmatrix} J^P_{\boldsymbol{\beta}\boldsymbol{\beta}} & J^P_{\boldsymbol{\beta}\Lambda} \\ J^P_{\Lambda\boldsymbol{\beta}} & J^P_{\Lambda\Lambda} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}_0 \\ \Lambda - \Lambda_0 \end{pmatrix},$$

where

$$\begin{aligned}
J^P_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) &= -P_0 \left\{ \frac{R_0(\int Q_0^{(1)} d\Lambda_0)(\int Q_0^{(1)} d\Lambda_0)^T}{(1+R_0)^2} + \frac{R_0(\int Q_0^{(2)} d\Lambda_0)}{1+R_0} \right\} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
J^P_{\boldsymbol{\beta}\Lambda}(\Lambda - \Lambda_0) &= -P_0 \left\{ \frac{R_0(\int Q_0^{(1)} d\Lambda_0) \int Q_0^{(0)} d(\Lambda - \Lambda_0)}{(1+R_0)^2} \right\} \\
J^P_{\Lambda\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)h &= -P_0 \left\{ \frac{R_0(\int Q_0^{(1)} d\Lambda_0)^T \int Q_0^{(0)} h d\Lambda_0}{(1+R_0)^2} \right\} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
J^P_{\Lambda\Lambda}(\Lambda - \Lambda_0)h &= -P_0 \left\{ \frac{R_0 \int Q_0^{(0)} h d\Lambda_0 \cdot \int Q_0^{(0)} h d(\Lambda - \Lambda_0)}{(1+R_0)^2} + \frac{R_0 \int Q_0^{(0)} h d(\Lambda - \Lambda_0)}{1+R_0} \right\},
\end{aligned}$$

where the functions with subscript zero are evaluated at the true parameter θ_0. Note that for $J^P_{\boldsymbol{\beta}\boldsymbol{\beta}}$ and $J^P_{\Lambda\Lambda}$, the second terms in the brackets have expectation zero, by the similar double expectation arguments as in (A.3). Since bounded linear operators with finite dimensional ranges are compact, we only need to show the compactness of $J^P_{\Lambda\boldsymbol{\beta}}$ and $J^P_{\Lambda\Lambda}$. That is to say, for a sequence of functions h_n in the unit ball, $J^P_{\Lambda\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)h_n$ and $J^P_{\Lambda\Lambda}(\Lambda - \Lambda_0)h_n$ have convergent subsequences. In fact, by (C1)-(C2) and the bounded variation properties of the functions involved, the convergent

subsequences can be selected using the Helly's lemma; thus, the operator J^P is compact.

We now show J is one-to-one. For $(b,h) \in \mathbb{R}^p \times BV[0,\tau]$, we need to show $J(b,h) = 0$ implies $b = 0$ and $h(t) = 0$. Similar to the arguments in Zeng and Lin (2006), some algebra gives

$$\begin{aligned} J(b,h) &= P_0 \left\{ \left(b^T \int_0^\tau \mathbf{Z}(dN - Ye^{\mathbf{Z}^T\boldsymbol{\beta}_0} d\Lambda_0) + \int_0^\tau h dN - \int_0^\tau Ye^{\mathbf{Z}^T\boldsymbol{\beta}_0} h d\Lambda_0 \right)^2 \right. \\ &\quad \left. + \frac{1}{R_0} \left\{ \frac{R_0}{1+R_0} b^T \int_0^\tau Q_0^{(1)} d\Lambda_0 + \frac{R_0}{1+R_0} \int_0^\tau Q_0^{(0)} h d\Lambda_0 \right)^2 \right\}. \end{aligned}$$

Comparing the expressions of J^C and J^P with V^C and V^P, we note that although the second Bartlett equality for the pairwise likelihood does not hold (Varin et al., 2011), the non-negativity of quadratic functions and R_0 indicate that, with probability one, the conditional score along the path $(\boldsymbol{\beta}_0 + b, \Lambda_0 + \varepsilon \int h d\Lambda_0)$

$$b^T \int_0^\tau \mathbf{Z}\{dN(s) - Y(s)e^{\mathbf{Z}^T\boldsymbol{\beta}_0} d\Lambda_0(s)\} + \int_0^\tau h(s) dN(s) - \int_0^\tau Y(s)e^{\mathbf{Z}^T\boldsymbol{\beta}_0} h(s) d\Lambda_0(s) = 0$$

By (C1) and (C3), considering the case of $N(\tau) = 0$ and $A + C \geqslant \tau$ and the case of $N(t) = I(t \geqslant t_0)$ for some $t_0 \in [0,\tau]$ and $A + C \geqslant \tau$, we obtain two equalities. Taking the difference, we have

$$\int_0^\tau \left(b^T\mathbf{Z} + h(s) \right) e^{\mathbf{Z}^T\boldsymbol{\beta}_0} d\Lambda_0(s) + b^T\mathbf{Z} + h(t_0) = 0.$$

The only solution to the above equations is trivial, thus

$$b^T\mathbf{Z} + h(t) = 0, \quad \forall t \in [0,\tau].$$

It follows from the identifiability condition (C2) that $b = 0$ and $h(t) = 0$.

With all four conditions verified, by Theorem 3.3.1 of van der Vaart and Wellner (1996), we have

$$n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow -J^{-1}\mathbb{G}_U,$$

where $\mathbb{G}_U$ is a mean-zero Gaussian process. Since linear maps preserve the Gaussian property, $\sqrt{n}(\hat{\theta} - \theta_0)$ also converge weakly to a mean-zero Gaussian process. In addition, the linear functional (3.7) converges weakly to a mean-zero Gaussian random variable with the variance estimator given by (3.8). The matrices $\hat{J}^C$ and $\hat{J}^P$ are given by

$$\begin{aligned}
\hat{J}^C &= -\frac{1}{n}\sum_{i=1}^{n} \partial U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}\,, \\
\hat{J}^P &= \frac{-1}{n(n-1)}\sum\sum_{i\neq j} \partial U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}\,.
\end{aligned}$$

The summand of the above matrices $\partial U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ and $\partial U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)$ take the forms

$$-\begin{pmatrix}
\mathbf{Z}_i^{\otimes 2} e^{\mathbf{Z}_i^T\boldsymbol{\beta}} \sum_{k=1}^m \lambda_k Y_i(w_k) & \mathbf{Z}_i e^{\mathbf{Z}_i^T\boldsymbol{\beta}} Y_i(w_1) & \cdots & \mathbf{Z}_i e^{\mathbf{Z}_i^T\boldsymbol{\beta}} Y_i(w_m) \\
\mathbf{Z}_i^T e^{\mathbf{Z}_i^T\boldsymbol{\beta}} Y_i(w_1) & I(X_i = w_1)\Delta_i/\lambda_1^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{Z}_i^T e^{\mathbf{Z}_i^T\boldsymbol{\beta}} Y_i(w_m) & 0 & \cdots & I(X_i = w_m)\Delta_i/\lambda_m^2
\end{pmatrix}$$

and

$$-R_{ij}\begin{pmatrix}
\frac{\left(\Lambda(Q_{ij}^{(1)})\right)^{\otimes 2}}{(1+R_{ij})^2} + \frac{\Lambda(Q_{ij}^{(2)})}{(1+R_{ij})} & \frac{Q_{ij}^{(0)}(w_1)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_1)}{(1+R_{ij})} & \cdots & \frac{Q_{ij}^{(0)}(w_m)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_m)}{(1+R_{ij})} \\
\left\{\frac{Q_{ij}^{(0)}(w_1)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_1)}{(1+R_{ij})}\right\}^T & \frac{\left(Q_{ij}^{(0)}(w_1)\right)^2}{(1+R_{ij})^2} & \cdots & \frac{Q_{ij}^{(0)}(w_1)Q_{ij}^{(0)}(w_m)}{(1+R_{ij})^2} \\
\vdots & \vdots & \ddots & \vdots \\
\left\{\frac{Q_{ij}^{(0)}(w_m)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_m)}{(1+R_{ij})}\right\}^T & \frac{Q_{ij}^{(0)}(w_1)Q_{ij}^{(0)}(w_m)}{(1+R_{ij})^2} & \cdots & \frac{\left(Q_{ij}^{(0)}(w_m)\right)^2}{(1+R_{ij})^2}
\end{pmatrix},$$

respectively, where

$$\Lambda(Q_{ij}^{(l)}) = \sum_{k=1}^{m} \lambda_k Q_{ij}^{(l)}(w_k).$$

The consistency of variance estimator (3.8) follows from the Glivenkon–Cantelli theorem and Remark A.3. □

A.2 Additional Simulation Results

This section contains some additional simulation results comparing the proposed method with the competitors.

Table A.1 displays the results under the same setup as in Section 3.3 with sample size $N = 200$. Note that the SEs for PLAC here are about twice those in Table 3.1 with $N = 800$.

PC			Conditional		LBML			PLAC				
		True	Bias	SE	Bias	SE	RE	Bias	SE	SEE	CP	RE
			Case 1: length-biased sampling; $A^* \sim$ Uniform(0,τ)									
50	$\hat{\beta}_1$	1	15	233	−58	160	1.87	17	184	177	94	1.60
	$\hat{\beta}_2$	1	3	211	−68	155	1.56	10	173	160	94	1.48
	$\hat{\Lambda}_{\tau_{30}}$	0.212	1	61	22	53	1.12	−1	56	54	93	1.18
	$\hat{\Lambda}_{\tau_{60}}$	0.546	5	121	56	101	1.10	0	112	111	94	1.18
80	$\hat{\beta}_1$	1	43	407	−88	198	3.57	48	257	238	95	2.46
	$\hat{\beta}_2$	1	22	360	−99	188	2.89	47	251	218	92	1.99
	$\hat{\Lambda}_{\tau_{30}}$	0.103	1	56	36	46	0.93	−4	47	42	87	1.41
	$\hat{\Lambda}_{\tau_{60}}$	0.329	4	136	60	94	1.50	−6	114	106	90	1.43
			Case 2: non-length-biased sampling; $A^* \sim$ Exponential(1)									
50	$\hat{\beta}_1$	1	8	218	−251	141	0.57	17	186	182	96	1.37
	$\hat{\beta}_2$	1	7	224	−246	148	0.61	20	191	183	94	1.36
	$\hat{\Lambda}_{\tau_{30}}$	0.207	−2	59	103	70	0.22	−4	57	53	92	1.06
	$\hat{\Lambda}_{\tau_{60}}$	0.538	−7	91	232	108	0.13	−9	89	90	94	1.02
80	$\hat{\beta}_1$	1	43	403	−388	170	0.92	59	303	264	92	1.73
	$\hat{\beta}_2$	1	61	400	−382	162	0.95	71	296	265	92	1.76
	$\hat{\Lambda}_{\tau_{30}}$	0.099	−5	46	107	65	0.14	−7	44	41	85	1.07
	$\hat{\Lambda}_{\tau_{60}}$	0.270	−10	86	215	105	0.13	−12	83	80	90	1.06

Table A.1: Summary of simulation with $N = 200$ and various censoring rates. PC: censoring percentage; True: true values; Bias, SE, SEE and CP: empirical bias ($\times 10^3$), standard error ($\times 10^3$), standard error estimate ($\times 10^3$) and 95% coverage probability; RE: asymptotic relative efficiency with respect to the conditional approach estimator (ratio of the mean squared errors). The estimate of $\hat{\Lambda}(t)$ is evaluated at the 30% and 60% percentiles (τ_{30} and τ_{60}) of the observed survival times.

We also tried the transformation approach to the Non-length-biaed case (Case 2), and compare its performance with the proposed method. Specifically, we first transformed the survival and truncation times using the distribution function of

Exponential(1), and then used the EM algorithm proposed by Qin et al. (2011) on the transformed data (Huang and Qin, 2012). The simulation with $N = 400$ and censoring rates 20%, 50% and 80% are reported in Table A.2.

PC			Conditional		LBML			PLAC				
		True	Bias	SE	Bias	SE	RE	Bias	SE	SEE	CP	RE
20	$\hat{\beta}_1$	1	9	117	−10	103	1.29	7	105	104	95	1.25
	$\hat{\beta}_2$	1	10	117	−8	104	1.26	8	107	104	94	1.19
	$\hat{\Lambda}_{\tau_{30}}$	0.298	0	43	4	39	1.19	−1	42	42	94	1.04
	$\hat{\Lambda}_{\tau_{60}}$	0.764	3	70	9	64	1.16	2	68	69	95	1.04
50	$\hat{\beta}_1$	1	9	152	−51	122	1.33	14	129	129	95	1.38
	$\hat{\beta}_2$	1	7	148	−55	124	1.21	11	130	129	94	1.29
	$\hat{\Lambda}_{\tau_{30}}$	0.207	−2	39	13	37	0.99	−3	38	38	94	1.04
	$\hat{\Lambda}_{\tau_{60}}$	0.538	−1	68	25	64	0.97	−3	67	64	93	1.03
80	$\hat{\beta}_1$	1	3	260	−121	158	1.71	26	191	181	93	1.83
	$\hat{\beta}_2$	1	2	262	−127	160	1.65	22	194	182	95	1.82
	$\hat{\Lambda}_{\tau_{30}}$	0.099	−1	34	26	36	0.57	−2	33	31	90	1.05
	$\hat{\Lambda}_{\tau_{60}}$	0.270	−5	60	46	59	0.64	−7	59	58	92	1.03

Table A.2: Summary of simulation using transformation approach suggested in Huang and Qin (2012). PC: censoring percentage; True: true values; Bias, SE, SEE and CP: empirical bias ($\times 10^3$), standard error ($\times 10^3$), standard error estimate ($\times 10^3$) and 95% coverage probability; RE: asymptotic relative efficiency with respect to the conditional approach estimator (ratio of the mean squared errors). The estimate of $\hat{\Lambda}(t)$ is evaluated at the 30% and 60% percentiles (τ_{30} and τ_{60}) of the observed survival times.

A.3 Additional Data Analysis Results

Graphical check of the uniform truncation assumption The significant test result in Section 3.4 for the uniform truncation assumption was confirmed by the graphical checking method proposed by Asgharian et al. (2006). When the assumption holds, the estimated survival functions of A and V should coincide. However, as shown in Figure A.1, the estimated survival curve of V is above that of A throughout, and the point-wise confidence intervals for A always stay beyond those of V. This means that in the RRI-CKD dataset, the uniform truncation assumption is violated.

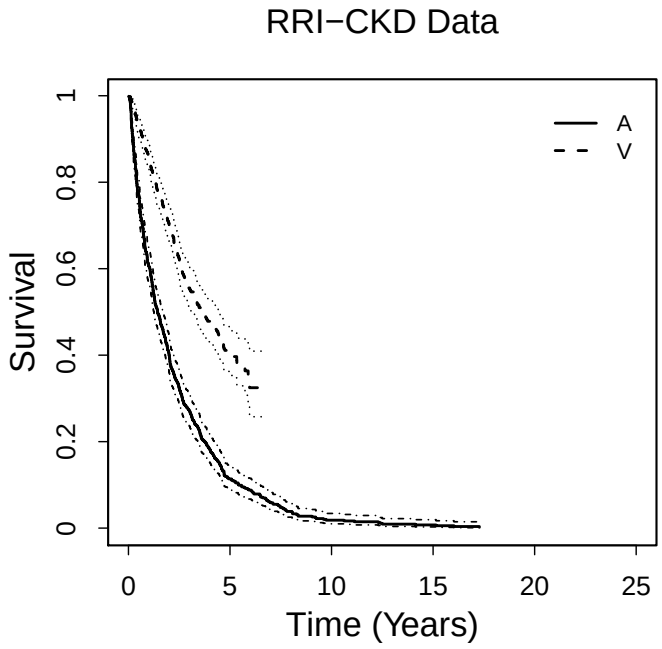

Figure A.1: Estimated survival curves for the truncation time A (solid) and the residual survival time V (dashed) of the RRI-CKD data. The 95% point-wise confidence intervals are shown as dashed or dotted lines around the estimates.

Regression coefficients estimates compared with the competitors A forest plot of the hazards ratios of the risk factors is shown in Figure A.2 to visualize the accuracy, precision and significance of the estimates. All coefficients have similar point estimates,

including the diabetes status and the CKD stage. However, the proposed estimator estimates all coefficients with improved precision indicated by narrower confidence intervals.

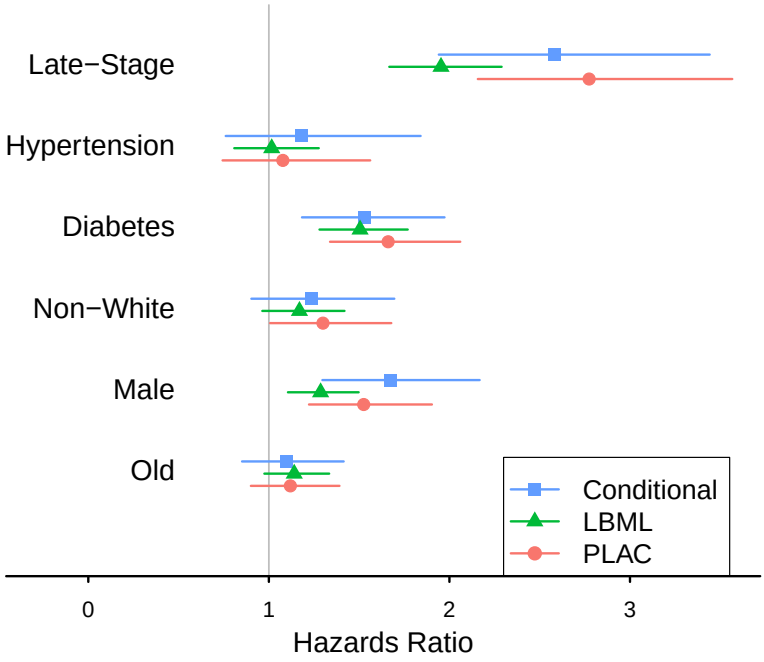

Figure A.2: Estimated hazards ratios of the covariates in the RRI-CKD data. The squares, triangles and dots represent the estimates using the conditional approach, the EM algorithm in Qin et al. (2011), and the proposed method (PLAC), respectively. The horizontal lines around the points represent the corresponding 95% confidence intervals.

Graphical check of the independence between A^* and Z^* We developed a graphical way to check the independence assumption between the *underlying* truncation time A^* and the covariates Z^*. To be specific, we estimate the unbiased truncation distribution G for each level of the covariate under investigation. The estimation of G follows closely to the inverse-probability weighted estimator proposed in Wang (1991) and Huang and Qin (2013). First, we calculate the right-truncation probability for observed A_i's using the survival probability calculated from the fitted Cox model with the conditional approach, and then use their reciprocals as the weights

to estimate G as a weighted empirical cumulative distribution function.

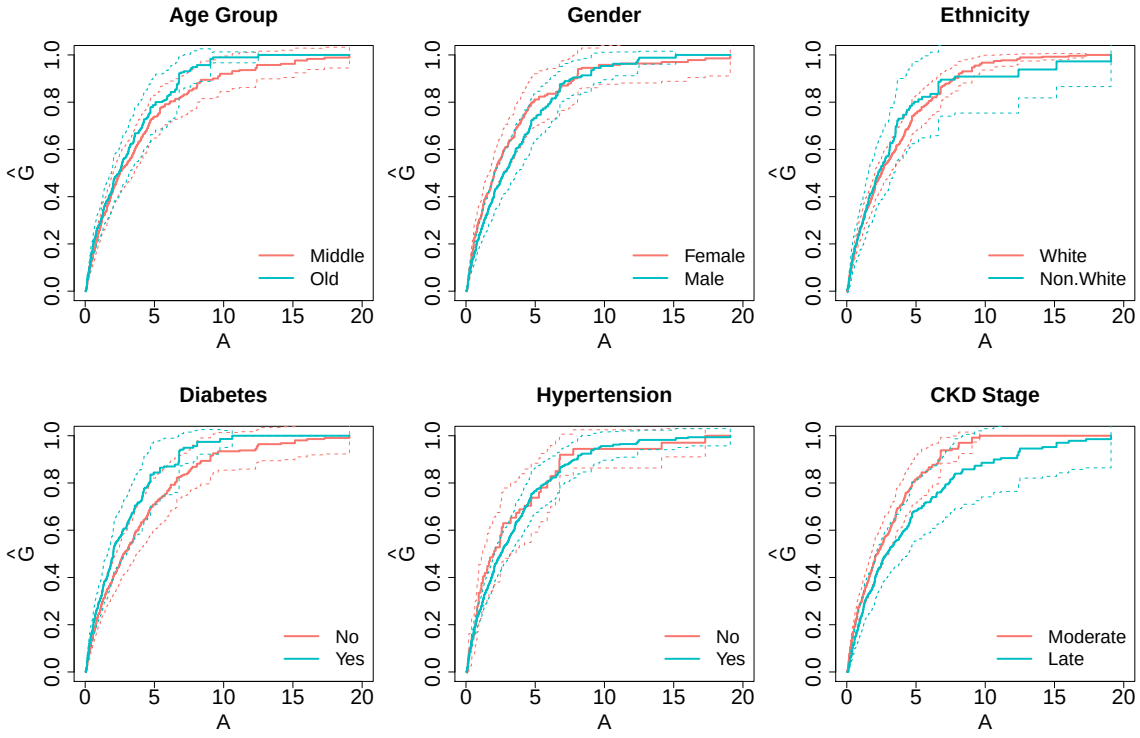

Figure A.3: Estimated $\hat{G}$ for each level of the covariates included in the RRI-CKD data analysis. The solid lines are the estimates, and the dashed line in the same color are the corresponding 95% point-wise confidence intervals.

APPENDIX B

Proofs, Additional Simulation and Data Analysis for the Second Project

B.1 Asymptotic Properties of the PLAC Estimator for Time-Dependent Covariates

The extension of the PLAC estimator to incorporating time-dependent covariates will not change the structures of most quantities involed in the proofs in A.1. Therefore, we establish the consistency and asymptotic normality of $(\hat{\boldsymbol{\beta}},\ \hat{\Lambda})$ utilizing similar steps as those in the last appendix with techniques from both empirical process and U-process theories. Note that the only change we need to take care is the covariates, which now dependes on time thus cannot be factored out from the cumulative hazards. Similar changes that are necessary for the modified pairwise likelihood (4.8) are trivial, and hence are omitted. Denote the scores corresponding to $\log \mathcal{L}_n^C$ and $\log \mathcal{L}_n^P$ as $U^C(\boldsymbol{\beta}, \boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^n U_i^C(\boldsymbol{\beta}, \Lambda)$ and $U^P(\boldsymbol{\beta}, \boldsymbol{\lambda}) = 2\{n(n-1)\}^{-1}\sum_{i<j} U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})$, where

(B.1)

$$U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \left\{ \Delta_i \mathbf{Z}_i^{\mathrm{T}}(t)(X_i) - \sum_{k=1}^m \lambda_k Y_i(w_k)\mathbf{Z}_i^{\mathrm{T}}(w_k) e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(w_k)}, \frac{\partial \ell_i^C}{\partial \lambda_1}, \ldots, \frac{\partial \ell_i^C}{\partial \lambda_m} \right\}^{\mathrm{T}},$$

(B.2)

$$U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda}) = -\frac{1}{1+R_{ij}^{-1}} \left\{ \sum_{k=1}^m \lambda_k Q_{ij}^{(1)T}(w_k), Q_{ij}^{(0)}(w_1), \ldots, Q_{ij}^{(0)}(w_m) \right\}^{\mathrm{T}},$$

and $\partial \ell_i^C / \partial \lambda_k = I(X_i = w_k)\{\Delta_i/\lambda_k - Y_i(w_k)\, e^{\boldsymbol{\beta}^{\mathrm{T}} \mathbf{Z}_i(w_k)}\}, k = 1, \ldots, m$. Let $(\boldsymbol{\beta}_0, \Lambda_0(\cdot))$ be the true parameter. The proofs are given under slightly modified regularity conditions as used in Appendix A.

(C1) The vector $\boldsymbol{\beta}_0$ lies in the interior of a compact set $B \subset \mathbb{R}^p$, and $\Lambda_0(\cdot)$ is continuously differentiable, strictly increasing on $[0, \tau]$, and satisfies $\Lambda_0(0) = 0$.

(C2) The covariates processes $\mathcal{Z}$ is uniformly bounded on $[0, \tau]$ with probability one. Moreover, if there exist a deterministic function $b_0(t)$ and a vector $b \in \mathbb{R}^p$, such that $b_0(t) + b^T \mathbf{Z}(t) = 0$ with probability one, then $b_0(t) = 0$ and $b = 0$.

(C3) With probability one, there exists a constant $\delta_1 > 0$ such that $\Pr(A^* < T^* \leqslant A^* + C | \mathcal{Z}, A^*, C) > \delta_1$, $\Pr(A + C \geqslant \tau \mid \mathcal{Z}) > \delta_1$, and that $\Pr(T \geqslant \tau \mid \mathcal{Z}) > \delta_1$.

(C4) Let $b \in \mathbb{R}^p$, and $h \in \mathrm{BV}[0, \tau]$, the space of all functions with bounded total variations on $[0, \tau]$, then the information operator corresponding to $\log \mathcal{L}_n^C$ evaluated at $(\boldsymbol{\beta}_0, \Lambda_0)$,

$$J_0^C(b,\, h) = \left(\lim_{n \to \infty} \partial U^C(\boldsymbol{\beta}, \Lambda) / \partial(\boldsymbol{\beta}, \Lambda) \Big|_{\boldsymbol{\beta} = \boldsymbol{\beta}_0,\, \Lambda = \Lambda_0} \right) (b,\, h)$$

is invertible.

As in Appendix A, if with probability one, there exists a constant $\delta_2 > 0$ such that $\Pr(A^* \geqslant T^* | \mathcal{Z}) > \delta_2$, then the PLAC estimator will be more effeicient than the corresponding conditional approach estimator, otherwise, it will reduce to the conditonal approach estimator. The identifiability of the parameter follows similar to the identifiability proof in Appendix A. We need to show the uniform convergence of the bivariate furncion classes in the pairwise likelihood and its derivatives through

bounding the bracketing numbers (entropies) of these function classes using the U-processes theory (De la Peña and Giné, 1999). To this end, we establish the following lemma on the $\sqrt{n}$-uniform convergence rate and asymptotic normality of the log-generalized odds ratio.

Lemma B.1. *Under Conditions (C1)-(C3), the log-generalized odds ratios process*

$$\mathcal{R} = \{(\mathcal{O}_i, \mathcal{O}_j) \mapsto r_{ij}(\boldsymbol{\beta}, \Lambda) : \mathcal{O}_i, \mathcal{O}_j \in \Omega, \boldsymbol{\beta} \in B, \Lambda \in \mathcal{H}_\Lambda\},$$

where $r_{ij}(\boldsymbol{\beta}, \Lambda) = \int_0^\tau (e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)})(I(t \leqslant A_i) - I(t \leqslant A_j))d\Lambda(t)$, *satisfies the uniform central limit theorem for U-processes:*

$$\sqrt{n}(\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda)) \rightsquigarrow \mathbb{G}_r,$$

where $\mathbb{G}_r$ *is a tight mean-zero Gaussian process.*

Proof. To establish the weak convergence, we first show that

$$\left\| \mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda) - \hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda) \right\|_{\boldsymbol{\beta},\Lambda} = o_p(n^{-1/2}),$$

where

$$\hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda) = \sum_{i=1}^{n} \mathrm{E}\left(\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda) | \mathcal{O}_i\right)$$

is the Hájek projection of $\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda)$ (van der Vaart, 2000), and $\|\cdot\|_{\boldsymbol{\beta},\Lambda}$ is the supreme norm over the parameter space.

By (C1)-(C3) and Fubini's theorem, we can interchange the order of the expectations and the intergrals. It can be verified that $P_0^2 r(\boldsymbol{\beta}, \Lambda) = 2\int_0^\tau \mathrm{Cov}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(t)}, I(t \leqslant A))d\Lambda(t)$. Moreover, since the pair $\mathcal{O}_i$ and $\mathcal{O}_j$ are i.i.d.,

$$\begin{aligned}
\mathrm{E}(r_{ij}(\boldsymbol{\beta}, \Lambda) \,|\, \mathcal{O}_i) &= \int_0^\tau \mathrm{E}\left\{ (e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)})(I(t \leqslant A_i) - I(t \leqslant A_j)) \,\middle|\, A_i, \mathcal{Z}_i \right\} d\Lambda(t) \\
&= \int_0^\tau \Big\{ e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} I(t \leqslant A_i) - \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}) I(t \leqslant A_i) \\
&\qquad - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} \mathrm{E}(I(t \leqslant A_i)) + \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} I(t \leqslant A_i)) \Big\} d\Lambda(t).
\end{aligned}$$

Thus the Hájek projection is

$$\begin{aligned}\hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta},\Lambda) &= \sum_{i=1}^{n}\mathrm{E}\left\{\binom{n}{2}^{-1}\sum_{j<k}r_{jk}(\boldsymbol{\beta},\Lambda) - P_0^2 r(\boldsymbol{\beta},\Lambda)\,\middle|\,\mathcal{O}_i\right\}\\ &= \frac{2}{n}\sum_{i=1}^{n}\int_0^\tau \Big\{e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}I(t\leqslant A_i) - \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)})I(t\leqslant A_i)\\ &\qquad - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}\mathrm{E}(I(t\leqslant A_i)) + \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}I(t\leqslant A_i))\\ &\qquad - 2\mathrm{Cov}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(t)}, I(t\leqslant A))\Big\}\,d\Lambda(t).\end{aligned}$$

Direct calculation gives

$$\tilde{\mathbb{U}}_{n,2} \equiv \mathbb{U}_{n,2}r(\boldsymbol{\beta},\Lambda) - P^2 r(\boldsymbol{\beta},\Lambda) - \hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta},\Lambda) = \frac{1}{\binom{n}{2}}\sum_{i<j}\tilde{\mathbb{U}}_{n,2}^{(i,j)}.$$

The summand of $\tilde{\mathbb{U}}_{n,2}$ is

$$\begin{aligned}\tilde{\mathbb{U}}_{n,2}^{(i,j)} &= \int_0^\tau \Big\{e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}I(t\leqslant A_i) - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)}I(t\leqslant A_i) - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}I(t\leqslant A_j) + e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)}I(t\leqslant A_j)\\ &\qquad - 2\mathrm{Cov}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(t)}, I(t\leqslant A))\\ &\qquad - \Big(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}I(t\leqslant A_i) - \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)})I(t\leqslant A_i) - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}\mathrm{E}(I(t\leqslant A_i))\\ &\qquad\quad + \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)}I(t\leqslant A_i)) - 2\mathrm{Cov}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(t)}, I(t\leqslant A))\Big)\\ &\qquad - \Big(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)}I(t\leqslant A_j) - \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)})I(t\leqslant A_j) - e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)}\mathrm{E}(I(t\leqslant A_j))\\ &\qquad\quad + \mathrm{E}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)}I(t\leqslant A_j)) - 2\mathrm{Cov}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(t)}, I(t\leqslant A))\Big)\Big\}\,d\Lambda(t)\\ &= -\int_0^\tau \Big\{(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} - \mathrm{E}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)})(I(t\leqslant A_j) - \mathrm{E}I(t\leqslant A_j))\\ &\qquad + (e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)} - \mathrm{E}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_j(t)})(I(t\leqslant A_i) - \mathrm{E}I(t\leqslant A_i))\Big\}\,d\Lambda(t),\end{aligned}$$

where the second equality holds by the definition of the covariance and the i.i.d. property of the observations. Therefore,

$$\begin{aligned}\tilde{\mathbb{U}}_{n,2} &= -\frac{1}{\binom{n}{2}}\int_0^\tau \sum_{i=1}^{n}\sum_{j=1}^{n}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} - \mathrm{E}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)})(I(t\leqslant A_j) - \mathrm{E}I(t\leqslant A_j))d\Lambda(t)\\ &\asymp -2\int_0^\tau \left\{\frac{1}{n}\sum_{i=1}^{n}(e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)} - \mathrm{E}e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(t)})\cdot\frac{1}{n}\sum_{j=1}^{n}(I(t\leqslant A_j) - \mathrm{E}I(t\leqslant A_j))\right\}d\Lambda(t),\end{aligned}$$

where $\asymp$ means asymptotically equivalent. By Donsker's theorem, the two summations in the brackets are both of order $O_p(n^{-1/2})$; thus, we can bound the supremum norm of $\tilde{\mathbb{U}}_{n,2}$ up to a constant:

$$\begin{aligned}\left\|\tilde{\mathbb{U}}_{n,2}\right\|_{\boldsymbol{\beta},\Lambda} &\lesssim \int_0^\tau \left\|n^{-1/2}\mathbb{G}_n e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}(t)}\right\|_{\boldsymbol{\beta}} \cdot \left\|n^{-1/2}(G_n - G)(t)\right\|_t d\Lambda(t) \\ &= o_p(n^{-1/2}),\end{aligned}$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$ denotes the empirical process and G_n denotes the empirical distribution function or A^*. Note here we use the fact that $\Lambda(\cdot)$ is a bounded monotonic function on $[0, \tau]$.

Therefore, $\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda) - P_0^2 r(\boldsymbol{\beta}, \Lambda)$ is equivalent to its projection $\hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda)$ up to a term of $o_p(n^{-1/2})$. The weak convergence of the projection $\hat{\mathbb{U}}_{n,2}r(\boldsymbol{\beta}, \Lambda)$ can be established using the VC theory, Theorem 2.7.5 of van der Vaart and Wellner (1996) and the Donsker's theorem for empirical processes. Combining these two facts leads to the weak convergence of $\mathbb{U}_{n,2}r(\boldsymbol{\beta}, \Lambda)$. □

Proof of Consistency. We follow the proof of Theorem III.1 in Appendix A closely. The main difference here is that we use Lemma B.1 instead of Lemma A.2 to control the entropy for the cumulative hazard functions separately, because of the time-dependent covariates. We can re-write the modified composite log-likelihood (4.4) and the composite score functions using the linear functional notations:

$$\begin{aligned}\ell_n^c(\boldsymbol{\beta}, \Lambda) &= \mathbb{P}_n \int_0^\tau \left\{(\log \Lambda\{s\} + \boldsymbol{\beta}^T \mathbf{Z}(s))dN(s) - Y(s)e^{\boldsymbol{\beta}^T \mathbf{Z}(s)} d\Lambda(s)\right\} \\ &\quad -\mathbb{U}_{n,2} \log(1 + R(\boldsymbol{\beta}, \Lambda)).\end{aligned}$$

Differentiating it with respect to $\boldsymbol{\beta}$ yields the composite score function for $\boldsymbol{\beta}$:

$$\begin{aligned}U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \Lambda) &= \mathbb{P}_n \int_0^\tau \mathbf{Z}(s) \left\{dN(s) - Y(s)e^{\boldsymbol{\beta}^T \mathbf{Z}(s)} d\Lambda(s)\right\} \\ &\quad -\mathbb{U}_{n,2} \left\{\frac{R(\boldsymbol{\beta}, \Lambda)}{1 + R(\boldsymbol{\beta}, \Lambda)} \int_0^\tau Q^{(1)}(s; \boldsymbol{\beta}) d\Lambda(s)\right\}.\end{aligned}$$

For $0 \leqslant t \leqslant \tau$ and $h(\cdot) = I(\cdot \leqslant t)$, define a perturbation of Λ by $d\Lambda_\varepsilon = (1 + \varepsilon h)d\Lambda$. The derivative of $\ell_n^c(\boldsymbol{\beta}, \Lambda_\varepsilon)$ with respect to ε evaluated at $\varepsilon = 0$ yields the composite score function for Λ in the direction of h:

$$\begin{aligned} U_\Lambda(\boldsymbol{\beta}, \Lambda)(h) &= \mathbb{P}_n \int_0^\tau h(s) \left\{ dN(s) - Y(s) e^{\boldsymbol{\beta}^T \mathbf{Z}(s)} d\Lambda(s) \right\} \\ &\quad - \mathbb{U}_{n,2} \left\{ \frac{R(\boldsymbol{\beta}, \Lambda)}{1 + R(\boldsymbol{\beta}, \Lambda)} \int_0^\tau Q^{(0)}(s; \boldsymbol{\beta}) h(s) d\Lambda(s) \right\}. \end{aligned}$$

We can write the composite score function

$$U(\boldsymbol{\beta}, \Lambda) = \begin{pmatrix} U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \Lambda) \\ U_\Lambda(\boldsymbol{\beta}, \Lambda)(h) \end{pmatrix}$$

as the summation of $U^C(\boldsymbol{\beta}, \Lambda)$ and $U^P(\boldsymbol{\beta}, \Lambda)$; the former is the conditional approach score function and has expectation zero. We can also show that $\mathrm{E}_0\{U^P(\boldsymbol{\beta}_0, \Lambda_0)\} = 0$, since the summand of U^P satisfies $\mathrm{E}_0\{U_{ij}^P(\boldsymbol{\beta}_0, \Lambda_0)\} = 0$, $1 \leqslant i < j \leqslant n$.

Since $\log \mathcal{L}_n^P$ is always negative, by the similar arguments as in Zeng and Lin (2006), we can show that the PLAC estimator has finite jump sizes, and that $\hat{\Lambda}(\tau)$ is bounded a.s. when $n \to \infty$. Because $\ell_n^c(\boldsymbol{\beta}, \Lambda)$ is maximized at the PLAC estimator $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$ over the whole model, it is certainly maximized along the parametric submodel $(\hat{\boldsymbol{\beta}}, \Lambda_\varepsilon)$ at $\varepsilon = 0$. Thus by the regularity conditions, the PLAC estimator is the solution to the composite score equations $U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \Lambda) = 0$ and $U_\Lambda(\boldsymbol{\beta}, \Lambda)(h) = 0$. Interchanging the summations and integrals in the second equation and rearranging the resulting terms, we have

(B.3)

$$\mathbb{P}_n \int_0^\tau h(s) dN(s) = \int_0^\tau h(s) \left\{ \mathbb{P}_n Y(s) e^{\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{Z}(s)} + \mathbb{U}_{n,2} \frac{R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})}{1 + R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})} Q^{(0)}(s; \hat{\boldsymbol{\beta}}) \right\} d\hat{\Lambda}(s).$$

Let

$$M_n(s; \hat{\boldsymbol{\beta}}, \hat{\Lambda}) = \mathbb{P}_n Y(s) e^{\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{Z}(s)} + \mathbb{U}_{n,2} \frac{R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})}{1 + R(\hat{\boldsymbol{\beta}}, \hat{\Lambda})} Q^{(0)}(s; \hat{\boldsymbol{\beta}})$$

denote the random function in the brackets. Replacing $h(s)$ with $h(s)/M_n(s;\hat{\boldsymbol{\beta}},\hat{\Lambda})$ on both sides of (B.3) yields the self-consistency solution of Λ:

$$\hat{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s;\hat{\boldsymbol{\beta}},\hat{\Lambda})}.$$

Inspired by the form of $\hat{\Lambda}$, we define another random step function

$$\tilde{\Lambda}(t) = \mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s;\boldsymbol{\beta}_0,\Lambda_0)}.$$

Let $M_0(s;\boldsymbol{\beta}_0,\Lambda_0) = P_0\{Y(s)e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}(s)}\}$. Since $\mathrm{E}_0\{U(\boldsymbol{\beta}_0,\Lambda_0)\} = 0$ and $\mathrm{E}_0\{U^P(\boldsymbol{\beta}_0,\Lambda_0)\} = 0$, the same algebra as we used to get $\hat{\Lambda}$ yields

$$\Lambda_0(t) = P_0 \int_0^t \frac{dN(s)}{M_0(s;\boldsymbol{\beta}_0,\Lambda_0)}.$$

Under the regularity conditions (C2)-(C3), by Lemma B.1, and the double expectation argument as we used in (A.3), $s \mapsto M_n(s;\boldsymbol{\beta}_0,\Lambda_0)$ is uniformly bounded away from zero and infinity, and is of uniformly bounded variation when n is sufficiently large. Therefore, by the Glivenko-Cantelli theorem and Remark A.3, we have

$$\|M_n(s;\boldsymbol{\beta}_0,\Lambda_0) - M_0(s;\boldsymbol{\beta}_0,\Lambda_0)\|_{L_\infty[0,\tau]} \overset{a.s.}{\to} 0$$

and

$$\left\|\mathbb{P}_n \int_0^t \frac{dN(s)}{M_n(s;\boldsymbol{\beta}_0,\Lambda_0)} - P_0 \int_0^t \frac{dN(s)}{M_n(s;\boldsymbol{\beta}_0,\Lambda_0)}\right\|_{L_\infty[0,\tau]} \overset{a.s.}{\to} 0,$$

where $\|\cdot\|_{L_\infty[0,\tau]}$ is the supreme norm over $[0,\tau]$. These results combined with the dominated convergence theorem yield

$$\left\|\tilde{\Lambda}(t) - \Lambda_0(t)\right\|_{L_\infty[0,\tau]} \overset{a.s.}{\to} 0.$$

By the definition of the PLAC estimator, the log-composite-likelihood evaluated at $(\hat{\beta},\hat{\Lambda})$ is greater than that evaluated at $(\beta_0,\tilde{\Lambda})$:

$$\mathbb{P}_n \int_0^\tau \left\{\log \frac{\hat{\Lambda}}{\tilde{\Lambda}}\{s\} + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathrm{T}}\mathbf{Z}(s)\right\} dN(s)$$

$$-\mathbb{P}_n \left\{\int_0^\tau Y(s)e^{\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{Z}(s)}d\hat{\Lambda}(s) - \int_0^\tau Y(s)e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}(s)}d\tilde{\Lambda}(s)\right\} - \mathbb{U}_{n,2}\log\frac{1+R(\hat{\boldsymbol{\beta}},\hat{\Lambda})}{1+R(\boldsymbol{\beta}_0,\tilde{\Lambda})} \quad \geqslant \quad 0.$$

By assumption, $\boldsymbol{\beta}$ is in a compact set, and that $\hat{\Lambda}(t) \leqslant \hat{\Lambda}(\tau)$ is bounded for $t \in [0, \tau]$ with probability one. Thus, by the Bolzano–Weierstrass theorem and the Helly's selection lemma, for every subsequence of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$, we can find a further subsequence (still denoted as $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$) along which $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^*$ for some $\boldsymbol{\beta}^*$ and $\hat{\Lambda}(t) \to \Lambda^*(t)$, $\forall t \in [0, \tau]$ for some monotone function Λ^* almost surely.

Since $\hat{\Lambda}(t)$ is absolutely continuous with respect to $\tilde{\Lambda}(t)$, let $\eta(t) = \lim_{n\to\infty} d\hat{\Lambda}/d\tilde{\Lambda}$ be a bounded measurable function, then $\Lambda^*(t) = \int_0^t \eta(s) d\Lambda_0(s)$ (Zeng and Lin, 2006). By (C1), $\Lambda^*(t)$ is absolutely continuous with respect to the Lebesgue measure and we denote its derivative as $\lambda^*(t)$. Thus we have the ratio $d\hat{\Lambda}/d\tilde{\Lambda}$ converges to $\eta(t) = \lambda^*(t)/\lambda_0(t)$. Again, by the Glivenko-Cantelli theorem, Lemma B.1, Remark A.3 and the dominant convergence theorem, the difference of the log-composite-likelihoods converges to

$$P_0 \int_0^\tau \left\{ \log \frac{\lambda^*}{\lambda_0}(s) + (\boldsymbol{\beta}^* - \boldsymbol{\beta}_0)^{\mathrm{T}} \mathbf{Z}(s) \right\} dN(s)$$
$$-P_0 \left\{ \int_0^\tau Y(s) e^{\boldsymbol{\beta}^{*T} \mathbf{Z}(s)} d\Lambda^*(s) - \int_0^\tau Y(s) e^{\boldsymbol{\beta}_0^{\mathrm{T}} \mathbf{Z}(s)} d\Lambda_0(s) \right\} - P_0 \log \frac{1 + R(\boldsymbol{\beta}^*, \Lambda^*)}{1 + R(\boldsymbol{\beta}_0, \Lambda_0)} \quad \geqslant \quad 0.$$

The left-hand side is the composite Kullback-Leibler divergence (Varin and Vidoni, 2005) of the density indexed by $(\boldsymbol{\beta}^*, \Lambda^*)$ from the true density, which by identifiability should be strictly negative unless $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$ and $\Lambda^* = \Lambda_0$. Since every subsequence of $(\hat{\boldsymbol{\beta}}, \hat{\Lambda})$ has a further subsequence converging to $(\boldsymbol{\beta}_0, \Lambda_0)$, we have the convergence of the entire sequence to the same limit. Finally, the uniform convergence of $\hat{\Lambda}(t)$ to $\Lambda_0(t)$ over $[0, \tau]$ follows from the continuity of Λ_0. □

Proof of Asymptotic Normality. Let θ denote the parameters $(\boldsymbol{\beta}, \Lambda)$. We proceed by checking the four conditions in Theorem 3.3.1 of van der Vaart and Wellner (1996). Note that $\sqrt{n}U(\theta_0)$ can be decomposed into $\sqrt{n}U^C(\theta_0) + \sqrt{n}U^P(\theta_0)$. Following the martingale theory, the first term converges weakly to a mean-zero Gaussian process

$\mathbb{G}_{U^C}$, and the linear functional

$$\sqrt{n}\left\{b_1^{\mathrm{T}} U_{\boldsymbol{\beta}}^C(\theta_0) + U_\Lambda^C(\theta_0)(h)\right\}$$

converges weakly to a mean-zero normal random variable with the variance that can be consistently estimated by $b^{\mathrm{T}}\hat{V}^C b$, where b is defined as in Section 2.3. For the second term, by Lemma B.1, the preservation theorem of Lipschitz functions and Theorem 5.3.1 of (De la Peña and Giné, 1999), it also converges weakly to a mean-zero Gaussian process $\mathbb{G}_{U^P}$, and the linear functional

$$\sqrt{n}\left\{b_1^{\mathrm{T}} U_{\boldsymbol{\beta}}^P(\theta_0) + U_\Lambda^P(\theta_0)(h)\right\}$$

converges weakly to a mean-zero normal random variable with the variance that can be consistently estimated by $b^{\mathrm{T}}\hat{V}^P b$. Note also that given $\{(A_i, \mathcal{Z}_i)\}_{i=1}^n$, $U^C(\theta_0)$ is a martingale, whereas $U^P(\theta_0)$ is a function of A_i and $\mathcal{Z}_i$ only, thus by the double expectation

$$\begin{aligned}
\mathrm{E}_0\{U^C(\theta_0)\cdot U^P(\theta_0)\} &= \mathrm{E}_0\left\{\mathrm{E}_0\left(U^C(\theta_0)\middle|\,(A_i,\mathcal{Z}_i), i=1,\ldots,n\right)\cdot U^P(\theta_0)\right\} \\
&= \mathrm{E}_0\{0\cdot U^P(\theta_0)\} = 0,
\end{aligned}$$

where $\cdot$ denotes the inner product of the underlying space. This indicates that the $U^C(\theta_0)$ and $U^P(\theta_0)$ are asymptotically independent (van der Vaart and Wellner, 1996, Example 1.4.6) at $\theta_0$that $\sqrt{n}U(\theta_0)$ converges weakly to a mean-zero Gaussian process $\mathbb{G}_U$. In addition, $\sqrt{n}\left\{b_1^{\mathrm{T}} U_{\boldsymbol{\beta}}(\theta_0) + U_\Lambda(\theta_0)(h)\right\}$ converges weakly to a mean-zero normal random variable with asymptotic variance that can be consistently estimated by $b^{\mathrm{T}}(\hat{V}^C + \hat{V}^P)\, b$. Therefore, the two stochastic conditions are satisfied by the consistency of $\hat{\theta}$, Lemma B.1 and Lemma 3.3.5 of van der Vaart and Wellner (1996). The fourth condition holds since $\hat{\theta}$ is a zero of $U(\theta)$ and that $u(\theta_0) \equiv \mathrm{E}_0 U(\theta_0) = 0$ by the arguments in the consistency proof.

To complete the proof, we only need to verify that the Fréchet-derivative of u at θ_0 exists and is continuous invertible. The Fréchet-differentiability can be check directly. For the continuous invertibility, note that the derivative $J \equiv \partial u(\theta)/\partial\theta|_{\theta=\theta_0}$ can be decomposed into J^C and J^P. By (C4) and the classic Cox model results, the first part is continuously invertible. Thus, it suffices to show J^P is a compact operator and that J is one-to-one by the Fredholm theory.

Following Example 3.3.10 of van der Vaart and Wellner (1996), we find the derivate J^P has the form

$$\begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}_0 \\ \Lambda - \Lambda_0 \end{pmatrix} \mapsto \begin{pmatrix} J^P_{\boldsymbol{\beta}\boldsymbol{\beta}} & J^P_{\boldsymbol{\beta}\Lambda} \\ J^P_{\Lambda\boldsymbol{\beta}} & J^P_{\Lambda\Lambda} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}_0 \\ \Lambda - \Lambda_0 \end{pmatrix},$$

where

$$\begin{aligned}
J^P_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) &= -P_0 \left\{ \frac{R_0(\int Q_0^{(1)} d\Lambda_0)(\int Q_0^{(1)} d\Lambda_0)^T}{(1+R_0)^2} + \frac{R_0(\int Q_0^{(2)} d\Lambda_0)}{1+R_0} \right\} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
J^P_{\boldsymbol{\beta}\Lambda}(\Lambda - \Lambda_0) &= -P_0 \left\{ \frac{R_0(\int Q_0^{(1)} d\Lambda_0) \int Q_0^{(0)} d(\Lambda - \Lambda_0)}{(1+R_0)^2} \right\} \\
J^P_{\Lambda\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)h &= -P_0 \left\{ \frac{R_0(\int Q_0^{(1)} d\Lambda_0)^T \int Q_0^{(0)} h d\Lambda_0}{(1+R_0)^2} \right\} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \\
J^P_{\Lambda\Lambda}(\Lambda - \Lambda_0)h &= -P_0 \left\{ \frac{R_0 \int Q_0^{(0)} h d\Lambda_0 \cdot \int Q_0^{(0)} h d(\Lambda - \Lambda_0)}{(1+R_0)^2} + \frac{R_0 \int Q_0^{(0)} h d(\Lambda - \Lambda_0)}{1+R_0} \right\},
\end{aligned}$$

where the functions with subscript zero are evaluated at the true parameter θ_0. Note that for $J^P_{\boldsymbol{\beta}\boldsymbol{\beta}}$ and $J^P_{\Lambda\Lambda}$, the second terms in the brackets have expectation zero, by the similar double expectation arguments as in (A.3). Since bounded linear operators with finite dimensional ranges are compact, we only need to show the compactness of $J^P_{\Lambda\boldsymbol{\beta}}$ and $J^P_{\Lambda\Lambda}$. That is to say, for a sequence of functions h_n in the unit ball, $J^P_{\Lambda\boldsymbol{\beta}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)h_n$ and $J^P_{\Lambda\Lambda}(\Lambda - \Lambda_0)h_n$ have convergent subsequences. In fact, by (C1)-(C2) and the bounded variation properties of the functions involved, the convergent subsequences can be selected using the Helly's lemma; thus, the operator J^P is

compact.

We now show J is one-to-one. For $(b,h) \in \mathbb{R}^p \times BV[0,\tau]$, we need to show $J(b,h) = 0$ implies $b = 0$ and $h(t) = 0$. Similar to the arguments in Zeng and Lin (2006), some algebra gives

$$\begin{aligned} J(b,h) &= P_0 \left\{ \left(b^{\mathrm{T}} \int_0^\tau \mathbf{Z}(dN - Ye^{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_0} d\Lambda_0) + \int_0^\tau h dN - \int_0^\tau Y e^{\mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_0} h d\Lambda_0 \right)^2 \right. \\ &\quad \left. + \frac{1}{R_0} \left\{ \frac{R_0}{1+R_0} b^{\mathrm{T}} \int_0^\tau Q_0^{(1)} d\Lambda_0 + \frac{R_0}{1+R_0} \int_0^\tau Q_0^{(0)} h d\Lambda_0 \right)^2 \right\}. \end{aligned}$$

Comparing the expressions of J^C and J^P with V^C and V^P, we note that although the second Bartlett equality for the pairwise likelihood does not hold (Varin et al., 2011), the non-negativity of quadratic functions and R_0 indicate that, with probability one, the conditional score along the path $(\boldsymbol{\beta}_0 + b, \Lambda_0 + \varepsilon \int h d\Lambda_0)$

$$b^{\mathrm{T}} \int_0^\tau \mathbf{Z}(s)\{dN(s) - Y(s)e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}} d\Lambda_0(s)\} + \int_0^\tau h(s) dN(s) - \int_0^\tau Y(s) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}(s)} h(s) d\Lambda_0(s) = 0$$

By (C1) and (C3), considering the case of $N(\tau) = 0$ and $A + C \geqslant \tau$ and the case of $N(t) = I(t \geqslant t_0)$ for some $t_0 \in [0,\tau]$ and $A + C \geqslant \tau$, we obtain two equalities. Taking the difference, we have

$$\int_0^\tau \left(b^{\mathrm{T}}\mathbf{Z}(s) + h(s) \right) e^{\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{Z}(s)} d\Lambda_0(s) + b^{\mathrm{T}}\mathbf{Z}(s) + h(t_0) = 0.$$

The only solution to the above equations is trivial, thus

$$b^{\mathrm{T}}\mathbf{Z}(t) + h(t) = 0, \quad \forall t \in [0,\tau].$$

It follows from the identifiability condition (C2) that $b = 0$ and $h(t) = 0$.

With all four conditions verified, by Theorem 3.3.1 of van der Vaart and Wellner (1996), we have

$$n^{1/2}(\hat{\theta} - \theta_0) \rightsquigarrow -J^{-1}\mathbb{G}_U,$$

where $\mathbb{G}_U$ is a mean-zero Gaussian process. Since linear maps preserve the Gaussian property, $\sqrt{n}(\hat{\theta} - \theta_0)$ also converge weakly to a mean-zero Gaussian process. In addition, the linear functional (3.7) converges weakly to a mean-zero Gaussian random variable with the variance estimator given by $\boldsymbol{\Sigma}$. The matrices $\hat{J}^C$ and $\hat{J}^P$ are given by

$$\begin{aligned}
\hat{J}^C &= -\frac{1}{n}\sum_{i=1}^{n} \partial U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}\,, \\
\hat{J}^P &= \frac{-1}{n(n-1)}\sum\sum_{i\neq j} \partial U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})\big|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}}\,.
\end{aligned}$$

The summand of the above matrices $\partial U_i^C(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})$ and $\partial U_{ij}^P(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})$ take the forms

$$-\begin{pmatrix}
\sum_{k=1}^m \lambda_k \mathbf{Z}_i^{\otimes 2}(w_k) e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(w_k)} Y_i(w_k) & \mathbf{Z}_i(w_1) e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(w_1)} Y_i(w_1) & \cdots & \mathbf{Z}_i(w_m) e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(w_m)} Y_i(w_m) \\
\mathbf{Z}_i^{\mathrm{T}}(w_1) e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(w_1)} Y_i(w_1) & I(X_i = w_1)\Delta_i/\lambda_1^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{Z}_i^{\mathrm{T}}(w_m) e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_i(w_m)} Y_i(w_m) & 0 & \cdots & I(X_i = w_m)\Delta_i/\lambda_m^2
\end{pmatrix}$$

and

$$-R_{ij}\begin{pmatrix}
\frac{\left(\Lambda(Q_{ij}^{(1)})\right)^{\otimes 2}}{(1+R_{ij})^2} + \frac{\Lambda(Q_{ij}^{(2)})}{(1+R_{ij})} & \frac{Q_{ij}^{(0)}(w_1)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_1)}{(1+R_{ij})} & \cdots & \frac{Q_{ij}^{(0)}(w_m)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_m)}{(1+R_{ij})} \\
\left\{\frac{Q_{ij}^{(0)}(w_1)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_1)}{(1+R_{ij})}\right\}^{\mathrm{T}} & \frac{\left(Q_{ij}^{(0)}(w_1)\right)^2}{(1+R_{ij})^2} & \cdots & \frac{Q_{ij}^{(0)}(w_1)Q_{ij}^{(0)}(w_m)}{(1+R_{ij})^2} \\
\vdots & \vdots & \ddots & \vdots \\
\left\{\frac{Q_{ij}^{(0)}(w_m)\Lambda(Q_{ij}^{(1)})}{(1+R_{ij})^2} + \frac{Q_{ij}^{(1)}(w_m)}{(1+R_{ij})}\right\}^{\mathrm{T}} & \frac{Q_{ij}^{(0)}(w_1)Q_{ij}^{(0)}(w_m)}{(1+R_{ij})^2} & \cdots & \frac{\left(Q_{ij}^{(0)}(w_m)\right)^2}{(1+R_{ij})^2}
\end{pmatrix},$$

respectively, where

$$\Lambda(Q_{ij}^{(l)}) = \sum_{k=1}^{m} \lambda_k Q_{ij}^{(l)}(w_k).$$

The consistency of variance estimator $\hat{\boldsymbol{\Sigma}}$ follows from the Glivenkon-Cantelli theorem and Remark A.3. □

B.2 Additional Simulation Results

We considered several additional simulation settings: different sample sizes and baseline hazards (Table B.1), Case 2 with various G (Table B.2), Case 3 with various G (Table B.3) and Case 1 with various F_ζ, the distribution of change point ζ (Table B.4).

G		Conditional		PLAC				
		Bias	SE	Bias	SE	SEE	CP	RE
Exp(1)	β_f	0.004	0.101	0.004	0.094	0.093	0.948	1.175
	β_v	0.000	0.104	0.000	0.103	0.102	0.950	1.019
	$\Lambda_0(\tau_{30})$	-0.002	0.050	-0.002	0.049	0.049	0.947	1.034
	$\Lambda_0(\tau_{70})$	0.001	0.106	0.001	0.103	0.105	0.962	1.059
Unif(0,100)	β_f	0.005	0.107	0.006	0.094	0.090	0.958	1.313
	β_v	0.002	0.107	0.000	0.106	0.103	0.948	1.016
	$\Lambda_0(\tau_{30})$	0.001	0.070	0.000	0.068	0.065	0.932	1.030
	$\Lambda_0(\tau_{70})$	0.003	0.148	0.003	0.143	0.137	0.937	1.070
Weib(3,3)	β_f	0.008	0.126	0.002	0.098	0.093	0.937	1.666
	β_v	0.007	0.112	0.007	0.112	0.106	0.923	1.008
	$\Lambda_0(\tau_{30})$	-0.012	0.276	-0.025	0.243	0.188	0.864	1.278
	$\Lambda_0(\tau_{70})$	0.003	0.379	-0.020	0.318	0.280	0.924	1.410
Bin(5,.2)	β_f	0.005	0.100	0.004	0.090	0.091	0.951	1.218
	β_v	0.004	0.104	0.004	0.103	0.101	0.953	1.029
	$\Lambda_0(\tau_{30})$	0.001	0.042	0.001	0.041	0.041	0.940	1.024
	$\Lambda_0(\tau_{70})$	0.004	0.129	0.003	0.123	0.119	0.939	1.105
DU(0:5)	β_f	0.002	0.103	0.004	0.094	0.091	0.940	1.200
	β_v	0.000	0.105	-0.001	0.103	0.101	0.949	1.032
	$\Lambda_0(\tau_{30})$	0.000	0.040	0.000	0.040	0.039	0.942	1.009
	$\Lambda_0(\tau_{70})$	0.005	0.126	0.005	0.121	0.118	0.938	1.084
Bin(5,.8)	β_f	0.008	0.141	0.009	0.102	0.100	0.942	1.911
	β_v	-0.008	0.109	-0.010	0.108	0.110	0.947	1.021
	$\Lambda_0(\tau_{30})$	-0.009	0.358	-0.024	0.300	0.274	0.917	1.414
	$\Lambda_0(\tau_{70})$	0.041	0.644	0.006	0.457	0.442	0.946	1.995

Table B.2: Summary of 1000 simulations in Case 2 with various G. N = 400 with no censoring (PC = 0%). True values $\beta_f = \beta_v = 1$ and $\lambda_0(t) = 2t$. τ_{30} and τ_{70} are the fixed 30% and 70% quantiles of observed event times for under each case. Exp(γ): exponential with rate γ; U(a, b): uniform (LBS) with limits a and b; WB(α, η): Weibull with shape α and scale η; Bin(n,p): binomial with n trails and probability p; DU$(a : b)$: discrete uniform distribution on integers from a to b. Bias: average difference between the estimates and the truth; SE: empirical standard error; SEE: estimated standard error; CP: 95% coverage probability; RE: relative efficiency (ratio of the mean squared errors).

N		Conditional		PLAC				
		Bias	SE	Bias	SE	SEE	CP	RE
		$\mathcal{Z} \perp A^*$; $Z_v(t) = I(t \geqslant \zeta)$, $\lambda_0(t) = 1$						
200	β_f	0.010	0.350	0.057	0.225	0.197	0.922	2.290
	β_v	0.011	0.385	0.019	0.241	0.232	0.942	2.529
	$\Lambda_0(\tau_{30})$	-0.007	0.153	-0.017	0.140	0.122	0.880	1.185
	$\Lambda_0(\tau_{70})$	-0.010	0.302	-0.015	0.266	0.245	0.915	1.286
800	β_f	0.005	0.158	0.017	0.098	0.097	0.948	2.503
	β_v	0.000	0.170	0.004	0.112	0.114	0.954	2.306
	$\Lambda_0(\tau_{30})$	-0.003	0.086	-0.005	0.087	0.068	0.911	0.979
	$\Lambda_0(\tau_{70})$	-0.001	0.149	-0.004	0.140	0.127	0.932	1.136
		$\mathcal{Z} \perp A^*$; $Z_v(t) = I(t \geqslant \zeta)$, $\lambda_0(t) = 2t$						
200	β_f	0.016	0.332	0.055	0.242	0.219	0.937	1.797
	β_v	0.021	0.368	0.043	0.255	0.237	0.932	2.033
	$\Lambda_0(\tau_{30})$	0.001	0.071	-0.003	0.063	0.054	0.882	1.289
	$\Lambda_0(\tau_{70})$	-0.002	0.174	-0.011	0.159	0.149	0.915	1.192
800	β_f	0.002	0.160	0.010	0.113	0.107	0.939	1.992
	β_v	0.007	0.169	0.010	0.121	0.115	0.947	1.961
	$\Lambda_0(\tau_{30})$	0.002	0.033	0.001	0.031	0.029	0.946	1.161
	$\Lambda_0(\tau_{70})$	-0.001	0.081	-0.003	0.076	0.075	0.936	1.148
		$\mathcal{Z} \not\perp A^*$; $Z_v(t) = I(t \geqslant A^* + \zeta_w)$, $\lambda_0(t) = 1$						
200	β_f	0.022	0.348	0.041	0.220	0.201	0.935	2.432
	β_v	-0.020	0.437	-0.045	0.424	0.380	0.942	1.052
	$\Lambda_0(\tau_{30})$	-0.012	0.203	-0.015	0.197	0.173	0.885	1.066
	$\Lambda_0(\tau_{70})$	0.038	0.463	0.031	0.413	0.390	0.928	1.260
800	β_f	0.008	0.162	0.016	0.106	0.101	0.937	2.279
	β_v	-0.001	0.194	-0.006	0.189	0.181	0.941	1.053
	$\Lambda_0(\tau_{30})$	-0.005	0.110	-0.006	0.106	0.096	0.925	1.072
	$\Lambda_0(\tau_{70})$	-0.004	0.233	-0.005	0.211	0.199	0.931	1.213
		$\mathcal{Z} \not\perp A^*$; $Z_v(t) = I(t \geqslant A^* + \zeta_w)$, $\lambda_0(t) = 2t$						
200	β_f	0.013	0.365	0.051	0.245	0.221	0.926	2.121
	β_v	0.008	0.490	-0.013	0.487	0.430	0.926	1.014
	$\Lambda_0(\tau_{30})$	-0.003	0.085	-0.005	0.082	0.073	0.877	1.086
	$\Lambda_0(\tau_{70})$	-0.009	0.225	-0.006	0.219	0.206	0.908	1.057
800	β_f	0.003	0.164	0.015	0.111	0.109	0.954	2.139
	β_v	0.011	0.206	0.007	0.203	0.203	0.945	1.027
	$\Lambda_0(\tau_{30})$	0.000	0.042	-0.001	0.041	0.038	0.931	1.040
	$\Lambda_0(\tau_{70})$	0.001	0.106	-0.001	0.102	0.104	0.953	1.063

Table B.1: Summary statistics from 1000 simulations with sample sizes $N = 200$ and 800, censoring rate (PC) 80% and baseline hazards $\lambda_0(t) = 1$ and $\lambda_0(t) = 2t$. True values $\beta_f = \beta_v = 1$. τ_{30} and τ_{70} are the fixed 30% and 70% quantiles of observed event times for under each case. Bias: average difference between the estimates and the truth; SE: empirical standard error; SEE: estimated standard error; CP: 95% coverage probability; RE: relative efficiency (ratio of the mean squared errors).

G		Conditional		PLAC				
		Bias	SE	Bias	SE	SEE	CP	RE
Exp(1)	β_f	0.004	0.101	0.004	0.092	0.091	0.945	1.193
	β_v	0.005	0.083	0.004	0.078	0.077	0.944	1.118
	$\Lambda_0(\tau_{30})$	0.001	0.046	0.000	0.046	0.046	0.942	1.013
	$\Lambda_0(\tau_{70})$	0.007	0.146	0.005	0.143	0.140	0.938	1.048
Unif(0,100)	β_f	0.006	0.101	0.005	0.088	0.089	0.953	1.302
	β_v	-0.003	0.083	-0.002	0.073	0.074	0.949	1.289
	$\Lambda_0(\tau_{30})$	0.000	0.058	0.000	0.056	0.058	0.953	1.064
	$\Lambda_0(\tau_{70})$	0.008	0.181	0.007	0.173	0.167	0.943	1.097
Weib(3,3)	β_f	0.007	0.119	0.003	0.086	0.090	0.956	1.926
	β_v	0.006	0.099	0.007	0.079	0.077	0.935	1.551
	$\Lambda_0(\tau_{30})$	-0.018	0.217	-0.029	0.196	0.164	0.891	1.205
	$\Lambda_0(\tau_{70})$	-0.005	0.385	-0.025	0.325	0.299	0.925	1.403
Bin(5,.2)	β_f	0.003	0.100	0.005	0.091	0.089	0.944	1.200
	β_v	0.006	0.084	0.008	0.076	0.072	0.938	1.224
	$\Lambda_0(\tau_{30})$	-0.001	0.044	-0.001	0.044	0.043	0.949	0.995
	$\Lambda_0(\tau_{70})$	0.005	0.163	0.007	0.156	0.153	0.952	1.097
DU(0:5)	β_f	0.003	0.100	0.004	0.092	0.090	0.945	1.170
	β_v	0.005	0.083	0.006	0.075	0.072	0.941	1.207
	$\Lambda_0(\tau_{30})$	1.341	0.129	1.341	0.123	0.121	0.000	1.001
	$\Lambda_0(\tau_{70})$	2.598	0.425	2.597	0.400	0.405	0.000	1.004
Bin(5,.8)	β_f	0.004	0.139	0.005	0.095	0.093	0.944	2.130
	β_v	0.005	0.108	0.008	0.082	0.077	0.938	1.688
	$\Lambda_0(\tau_{30})$	0.022	0.334	-0.002	0.261	0.243	0.924	1.640
	$\Lambda_0(\tau_{70})$	0.082	0.705	0.046	0.485	0.463	0.944	2.122

Table B.3: Summary of 1000 simulations in Case 3 with various G. N = 400 with no censoring (PC = 0%). True values $\beta_f = \beta_v = 1$ and $\lambda_0(t) = 2t$. τ_{30} and τ_{70} are the fixed 30% and 70% quantiles of observed event times for under each case. Exp(γ): exponential with rate γ; U(a, b): uniform (LBS) with limits a and b; WB(α, η): Weibull with shape α and scale η; Bin(n,p): binomial with n trails and probability p; DU($a : b$): discrete uniform distribution on integers from a to b. Bias: average difference between the estimates and the truth; SE: empirical standard error; SEE: estimated standard error; CP: 95% coverage probability; RE: relative efficiency (ratio of the mean squared errors).

F_ζ		Conditional		PLAC				
		Bias	SE	Bias	SE	SEE	CP	RE
WB(.5,.5)	β_f	0.004	0.103	0.005	0.091	0.090	0.945	1.304
	β_v	0.001	0.112	0.003	0.101	0.100	0.945	1.249
	$\Lambda_0(\tau_{30})$	-0.001	0.052	-0.002	0.050	0.049	0.937	1.062
	$\Lambda_0(\tau_{70})$	0.002	0.130	0.000	0.124	0.119	0.933	1.095
WB(1,1)	β_f	0.004	0.105	0.004	0.092	0.090	0.944	1.295
	β_v	-0.001	0.108	0.000	0.097	0.093	0.938	1.217
	$\Lambda_0(\tau_{30})$	-0.001	0.056	-0.002	0.055	0.053	0.941	1.052
	$\Lambda_0(\tau_{70})$	0.003	0.128	0.001	0.124	0.115	0.932	1.070
WB(1.5,1.5)	β_f	0.005	0.108	0.006	0.096	0.090	0.934	1.243
	β_v	0.001	0.108	0.002	0.096	0.094	0.945	1.245
	$\Lambda_0(\tau_{30})$	-0.001	0.062	-0.002	0.061	0.058	0.928	1.042
	$\Lambda_0(\tau_{70})$	0.003	0.127	0.002	0.122	0.118	0.933	1.080
WB(2,2)	β_f	0.003	0.106	0.004	0.093	0.091	0.941	1.310
	β_v	0.001	0.116	0.002	0.105	0.105	0.944	1.217
	$\Lambda_0(\tau_{30})$	-0.001	0.066	-0.002	0.065	0.064	0.937	1.025
	$\Lambda_0(\tau_{70})$	0.004	0.135	0.002	0.131	0.124	0.926	1.071

Table B.4: Summary of 1000 simulations in Case 1 with various F_ζ, the distribution of the change point. N = 400 with no censoring (PC = 0%). True values $\beta_f = \beta_v = 1$ and $\lambda_0(t) = 2t$. τ_{30} and τ_{70} are the fixed 30% and 70% quantiles of observed event times for under each case. WB(α, η): Weibull with shape α and scale η. Bias: average difference between the estimates and the truth; SE: empirical standard error; SEE: estimated standard error; CP: 95% coverage probability; RE: relative efficiency (ratio of the mean squared errors).

B.3 Additional Data Analysis Results

State	Patients	Deaths	Censoring (%)
CA	2912	841	71.1
TX	1782	498	72.1
NY	1597	494	69.1
PA	1203	435	63.8
FL	1040	316	69.6
IL	971	341	64.9
MI	783	285	63.6
OH	769	262	65.9
VA	653	261	60.0
NC	549	161	70.7
TN	516	167	67.6
AL	495	217	56.2
GA	483	142	70.6
MN	482	147	69.5
NJ	476	169	64.5
AZ	432	118	72.7
MA	432	151	65.0
MO	430	118	72.6
MD	376	121	67.8
WI	347	112	67.7
LA	329	108	67.2
IN	322	99	69.3
CO	284	100	64.8
WA	278	76	72.7
DC	254	73	71.3
OR	209	66	68.4
KY	183	63	65.6
OK	140	56	60.0
SC	140	42	70.0
AR	137	50	63.5
UT	137	30	78.1
NE	136	42	69.1
CT	119	32	73.1
IA	119	42	64.7
KS	103	28	72.8
NM	84	27	67.9
NV	81	25	69.1
ME	69	23	66.7
ND	66	19	71.2
WV	62	29	53.2

Table B.5: Number of patients, number of deaths, and censoring rates for the included 40 US states in th e OPTN/UNOS data.

APPENDIX C

Algorithm, Simulation Setup and Data Analysis Results for the Third Project

C.1 An Alternating Direction Method of Multiplier

We first describe the alternating direction method of multiplier (ADMM) to minimize the objective function (5.3). To fix ideas, let $y_i = (y_{i1}, \ldots, y_{in_i})^{\mathrm{T}}$, and denote the vector of all observations by $\mathbf{y} = (y_1^{\mathrm{T}}, \ldots, y_m^{\mathrm{T}})^{\mathrm{T}}$ and the corresponding basis expansion coefficients vector $\boldsymbol{\beta} = (\beta_1^{\mathrm{T}}, \ldots, \beta_m^{\mathrm{T}})^{\mathrm{T}}$. Let $S_i = (s(t_{i1}), \ldots, s(t_{in_i}))$, and $\mathbf{S} = \mathrm{bdiag}(S_1^{\mathrm{T}}, \ldots, S_m^{\mathrm{T}})$, where $\mathrm{bdiag}(\cdot)$ constructs a block diagonal matrix with the matrices inside the parentheses. Let $A_{ij} = (e_i - e_j)^{\mathrm{T}} \otimes I_p$, where e_i is an m-vector such that the i-th element is one and the rest are zeros, $\otimes$ is the Kronecker product, and I_p is the $p \times p$ identity matrix. Denote the set of pairs with non-zero weights by $\mathcal{H} = \{(i,j) : w_{ij} > 0\}$. Let $\mathbf{A}$ be the $Hp \times mp$ matrix stacking the matrices A_h over the pairs $h \in \mathcal{H}$, and $\mathbf{Q} = I_H \otimes Q^{\mathrm{T}}$, where H is the cardinality of $\mathcal{H}$. Then the augmented Lagrangian for the constraint optimization problem (5.3) is

$$\begin{aligned} L_\upsilon(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda}) =& \frac{1}{2} \left\| \mathbf{y} - \mathbf{S}\boldsymbol{\beta} \right\|_2^2 + \gamma \sum_{h \in \mathcal{H}} w_h \left\| \alpha_h \right\| \\ & + \langle \boldsymbol{\lambda}, \boldsymbol{\alpha} - \mathbf{Q}\mathbf{A}\boldsymbol{\beta} \rangle + \frac{\upsilon}{2} \left\| \boldsymbol{\alpha} - \mathbf{Q}\mathbf{A}\boldsymbol{\beta} \right\|_2^2, \end{aligned} \tag{C.1}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$ are the vectors obtained by stacking Lagrange multipliers λ_h and α_h over $\mathcal{H}$, respectively, and $\upsilon \geqslant 0$ is the penalty parameter for the augmented term.

To solve for the minimizer of (5.3), we can minimize (C.1) over $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\lambda})$. The three groups of variables are update iteratively (Boyd et al., 2011; Chi and Lange, 2015). At Step $(r+1)$, $r = 0, 1, \ldots$, the updating step for β amounts to minimizing

$$f(\beta) = \frac{1}{2} \|\mathbf{y} - \mathbf{S}\boldsymbol{\beta}\|_2^2 + \frac{1}{2} \|\boldsymbol{\alpha}_* - \mathbf{Q}\mathbf{A}\boldsymbol{\beta}\|_2^2$$

where $\boldsymbol{\alpha}_* = \boldsymbol{\alpha} + \upsilon^{-1}\boldsymbol{\lambda}$. Taking derivative of f with respect to $\boldsymbol{\beta}$, we have

$$\boldsymbol{\beta}^{(r+1)} = (\mathbf{S}^{\mathrm{T}}\mathbf{S} + \upsilon\mathbf{A}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{A})^{-1}(\mathbf{S}^{\mathrm{T}}\mathbf{y} + \upsilon\mathbf{A}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\boldsymbol{\alpha}_*^{(r)}). \tag{C.2}$$

The update of $\boldsymbol{\alpha}$ can be accomplished by the proximal minimization (Parikh and Boyd, 2014). First we note that (C.1) is separable in α_h, $h \in \mathcal{H}$. For given $h = (h_1, h_2)$, let $\sigma_h = \gamma w_h / \upsilon$, then the minimizor is determined by the proximal map of the norm $\|\cdot\|$:

$$\begin{aligned} \alpha_h^{(r+1)} &= \mathrm{argmin}_{\alpha_h} \frac{\gamma w_h}{\upsilon} \|\alpha_h\| + \frac{1}{2} \|\alpha_h - \{Q^{\mathrm{T}}(\beta_{h_1} - \beta_{h_2}) - \lambda_h/\upsilon\}\|_2^2 \\ &= \mathrm{prox}_{\sigma_h\|\cdot\|}\{Q^{\mathrm{T}}(\beta_{h_1} - \beta_{h_2}) - \lambda_h/\upsilon\}. \end{aligned} \tag{C.3}$$

Since the norm used in the pairwise differences is the L_1-norm, the updating step (C.3) is equivalent to element-wise soft thresholding (Boyd et al., 2011).

Finally, the Lagrange multipliers are updated by

$$\lambda_h^{(r+1)} = \lambda_h^{(r)} + \upsilon\{\alpha_h^{(r+1)} - Q^{\mathrm{T}}(\beta_{h_1}^{(r+1)} - \beta_{h_2}^{(r+1)})\},\ h \in \mathcal{H}. \tag{C.4}$$

In summary, the ADMM updating algorithm proceeds as follows: Start with some initial values for $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$; use (C.2)–(C.4) to update the parameters and the multipliers until convergence criteria are met. To check convergence, we follow the suggestions by Boyd et al. (2011). It is known that ADMM usually converges slower than the Newton-type optimization algorithms, but since the computation cost for

each iteration is cheap, the computation time to get a solution path is still reasonable for datasets with moderate sample sizes.

Convergence of the ADMM algorithm for convex clustering is guaranteed for any $\nu > 0$, and the different magnitudes of ν only change the weights on the proximal or dual residuals in the convergence criteria (Boyd et al., 2011; Chi and Lange, 2015). The convergence property of our modified clustering algorithm is beyond the scope of the current chapter and warrants further research. In our simulations and data analysis, we did not observe any non-convergent ADMM iterations as long as the maximum number of iterations were chosen large enough.

The only change we need to make in the ADMM algorithm to minimize (5.5) is in the first step, which is now decomposed into two steps:

$$\mathbf{u}^{(r+1)} = (\mathbf{S}^{\mathrm{T}}\mathbf{S} + \mathbf{G}^{-1})^{-1}(\mathbf{S}^{\mathrm{T}}(\mathbf{y} - \mathbf{S}\boldsymbol{\beta}^{(r)})), \tag{C.5}$$

$$\boldsymbol{\beta}^{(r+1)} = (\mathbf{S}^{\mathrm{T}}\mathbf{S} + \upsilon\mathbf{A}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{Q}\mathbf{A})^{-1}(\mathbf{S}^{\mathrm{T}}(\mathbf{y} - \mathbf{S}\mathbf{u}^{(r+1)}) + \upsilon\mathbf{A}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\boldsymbol{\alpha}_{*}^{(r)}), \tag{C.6}$$

where $\mathbf{u} = (u_1^{\mathrm{T}}, \ldots, u_m^{\mathrm{T}})^{\mathrm{T}}$ are the stacked vector of random effects, and $\mathbf{G} = I_m \otimes G$.

C.2 Simulation Setups

True cluster centers and example trajectories Figure C.1 gives the profiles of the true cluster centers as used in our simulation studies. Figure C.2 and Figure C.3 provide examples under various scenarios of σ_u and σ_e for Case 1 and 2 in the first set of simulations. The example of different sparsity and sample sizes are given in Figure C.4.

Parameter setup used in the simulations For the functional clustering model by James and Sugar (2003), the dimension of the natural splines was $q = 3$, where evenly spaced knots were used; the dimension for space that the mean coefficients were assumed to lie within were $h = 2$; the covariance of the random effects will have rank constraint $p = 5$. For the distance-based clustering method by Peng and Müller (2008), we used the default setting from R package `funcy`. As for the mixture mixed effect model, we followed the example given in R package `fdapace`. Specifically, we used generalized cross validation to choose the bandwidth of the mean and covariance functions. Bayesian information criterion was used to select the number of principal components, and the threshold for proportion of variance explained was 99%.

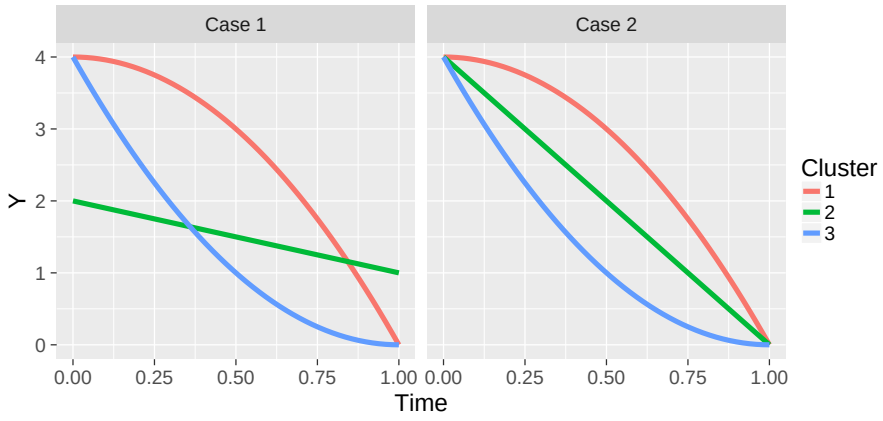

Figure C.1: The profiles of the true cluster centers used in the simulation.

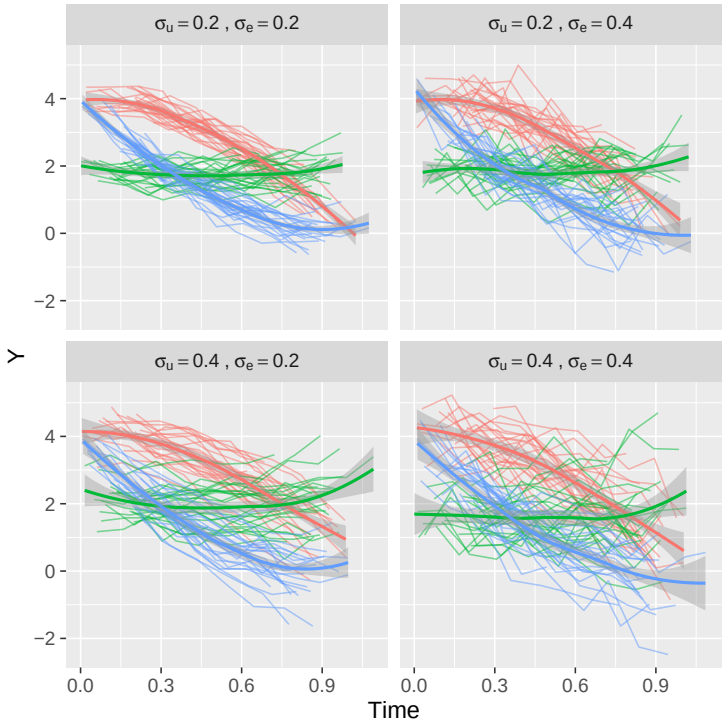

Figure C.2: Example trajectories for Simulation I, Case 1.

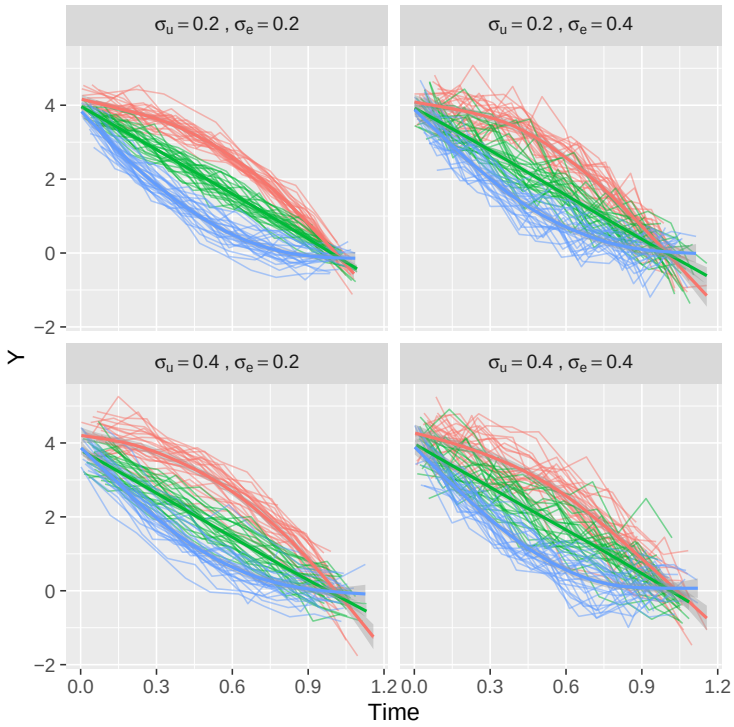

Figure C.3: Example trajectories for Simulation I, Case 2.

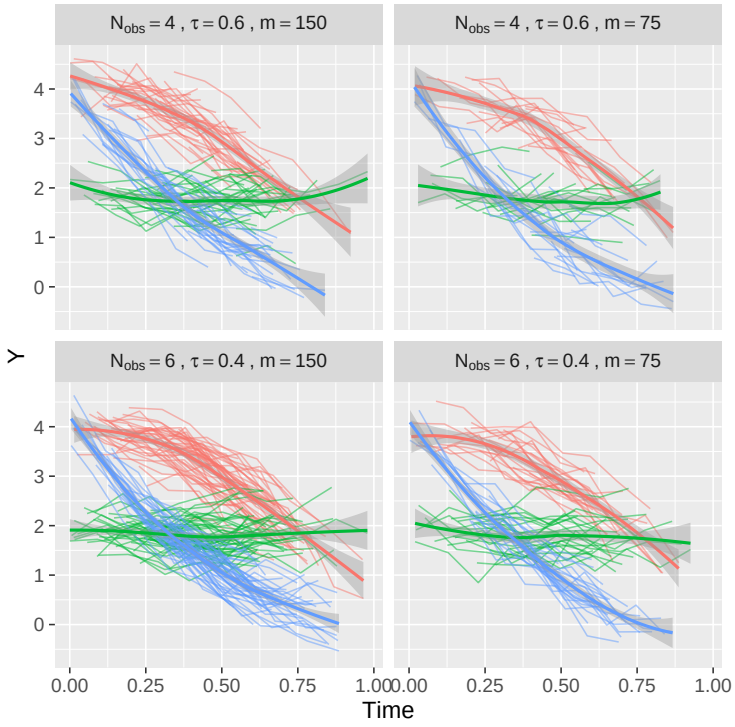

Figure C.4: Example trajectories for Simulation II.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aerts, M., Molenberghs, G., Ryan, L. M., and Geys, H. (2002). *Topics in modelling of clustered data*. CRC Press.

Aldwin, C. M., Spiro III, A., Levenson, M. R., and Cupertino, A. P. (2001). Longitudinal findings from the normative aging study: Iii. personality, individual health trajectories, and mortality. *Psychology and Aging*, 16(3):450.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Asgharian, M., M'Lan, C. E., and Wolfson, D. B. (2002). Length-biased sampling with right censoring: An unconditional approach. *Journal of the American Statistical Association*, 97(457):201–209.

Asgharian, M., Wolfson, C., and Wolfson, D. B. (2014). Analysis of biased survival data: The canadian study of health and aging and beyond. *Statistics in Action: A Canadian Outlook*, pages 193–208.

Asgharian, M., Wolfson, D. B., and Zhang, X. (2006). Checking stationarity of the incidence rate using prevalent cohort survival data. *Statistics in medicine*, 25(10):1751–1767.

Bates, D. M. and DebRoy, S. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, 91(1):1–17.

Bell, B., Rose, C. L., and Damon, A. (1972). The normative aging study: an interdisciplinary and longitudinal study of health and aging. *The International Journal of Aging and Human Development*, 3(1):5–17.

Bellio, R. and Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, 5(3):217–227.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236.

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F. (2006). Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics & Data Analysis*, 51(2):587–600.

Blumenthal, S. (1967). Proportional sampling in life length studies. *Technometrics*, 9(2):205–218.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Bruckers, L., Molenberghs, G., Drinkenburg, P., and Geys, H. (2016). A clustering algorithm for multivariate longitudinal data. *Journal of biopharmaceutical statistics*, 26(4):725–741.

Chi, E. C., Allen, G. I., and Baraniuk, R. G. (2016). Convex biclustering. *Biometrics*, (Accepted).

Chi, E. C. and Lange, K. (2015). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013.

Chiou, J.-M. and Li, P.-L. (2007). Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):679–699.

Cox, D. (1969). Some sampling problems in technology. In Johnson, N. L. and Smith Jr., H., editors, *New Developments in Survey Sampling,*, pages 506–527. Wiley, New York.

Cox, D. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 91(3):729–737.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 34(2):187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Crowley, J. and Hu, M. (1977). Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, 72(357):27–36.

De la Peña, V. and Giné, E. (1999). *Decoupling: from dependence to independence*. Springer.

De Leon, A. (2005). Pairwise likelihood approach to grouped continuous model and its extension. *Statistics & probability letters*, 75(1):49–57.

De Leon, A., Alexander, R., and Carrièger, K. (2007). General mixed-data model: Extension of general location and grouped continuous models. *Canadian Journal of Statistics*, 35(4):533–548.

de Uña-álvarez, J. (2004). Nonparametric estimation under length-biased sampling and type I censoring: a moment based approach. *Annals of the Institute of Statistical Mathematics*, 56(4):667–681.

Eiter, T. and Mannila, H. (1994). Computing discrete Fréchet distance. Technical report, Citeseer.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.

Gao, X. and Song, P. X.-K. (2011). Composite likelihood EM algorithm with applications to multivariate hidden markov model. *Statistica Sinica*, 21:165–185.

Genolini, C. and Falissard, B. (2010). KmL: k-means for longitudinal data. *Computational Statistics*, 25(2):317–328.

Genolini, C., Pingault, J., Driss, T., Côté, S., Tremblay, R. E., Vitaro, F., Arnaud, C., and Falissard, B. (2013). KmL3D: a non-parametric algorithm for clustering joint trajectories. *Computer methods and programs in biomedicine*, 109(1):104–111.

Heaf, J. G., Løkkegaard, H., and Madsen, M. (2002). Initial survival advantage of peritoneal dialysis relative to haemodialysis. *Nephrology Dialysis Transplantation*, 17(1):112–117.

Hocking, T. D., Joulin, A., Bach, F., and Vert, J.-P. (2011). Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*.

Huang, C.-Y. and Qin, J. (2011). Nonparametric estimation for length-biased and right-censored data. *Biometrika*, 98(1):177–186.

Huang, C.-y. and Qin, J. (2012). Composite partial likelihood estimation under length-biased sampling, with application to a prevalent cohort study of dementia. *Journal of the American Statistical Association*, 107(499):946–957.

Huang, C.-Y. and Qin, J. (2013). Semiparametric estimation for the additive hazards model with left-truncated and right-censored data. *Biometrika*, 100(4):877–888.

Huang, C.-Y., Qin, J., and Follmann, D. A. (2012). A maximum pseudo-profile likelihood estimator for the cox model under length-biased sampling. *Biometrika*, 99(1):199–210.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Jacques, J. and Preda, C. (2014). Functional data clustering: a survey. *Advances in Data Analysis and Classification*, 8(3):231–255.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408.

Jung, S.-H. (1999). Rank tests for matched survival data. *Lifetime Data Analysis*, 5(1):67–79.

Jung, Y., Park, H., Du, D.-Z., and Drake, B. L. (2003). A decision criterion for the optimal number of clusters in hierarchical clustering. *Journal of Global Optimization*, 25(1):91–111.

Kalbfleisch, J. and Lawless, J. (1991). Regression models for right truncated data with applications to AIDS incubation times and reporting lags. *Statistica Sinica*, 1:19–32.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. John Wiley & Sons, Inc., New York.

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, 69(1):19–27.

Korevaar, J. C., Feith, G., Dekker, F. W., van Manen, J. G., Boeschoten, E. W., Bossuyt, P. M., and T KREDIET, R. (2003). Effect of starting with hemodialysis compared with peritoneal dialysis in patients new on dialysis treatment: a randomized controlled trial. *Kidney international*, 64(6):2222–2228.

Liang, G. and Yu, B. (2003). Maximum pseudo likelihood estimation in network tomography. *Signal Processing, IEEE Transactions on*, 51(8):2043–2053.

Liang, K.-Y. and Qin, J. (2000). Regression analysis under non-standard situations: a pairwise pseudolikelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):773–786.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, 80(1):221–39.

Lindsten, F., Ohlsson, H., and Ljung, L. (2011). Just relax and come clustering!: A convexification of k-means clustering. Technical report, Linköpings University.

Liu, H., Ning, J., Qin, J., and Shen, Y. (2016). Semiparametric maximum likelihood inference for truncated or biased-sampling data. *Statistica Sinica*, 26(3):1087–1115.

Ma, S. and Huang, J. (2016a). A concave pairwise fusion approach to subgroup analysis. *Journal of the American Statistical Association*, (Accepted).

Ma, S. and Huang, J. (2016b). Estimating subgroup-specific treatment effects via concave fusion. *arXiv preprint arXiv:1607.03717*.

Mandel, M. (2007). Censoring and truncation–highlighting the differences. *The American Statistician*, 61(4):321–324.

Mandel, M. and Betensky, R. A. (2007). Testing goodness of fit of a uniform truncation model. *Biometrics*, 63(2):405–412.

Markides, K. S. (2007). *Encyclopedia of health and aging*. Sage Publications.

McDonald, S. P. and Craig, J. C. (2004). Long-term survival of children with end-stage renal disease. *New England Journal of Medicine*, 350(26):2654–2662.

McDonald, S. P. and Russ, G. R. (2002). Survival of recipients of cadaveric kidney transplants compared with those receiving dialysis treatment in australia and new zealand, 1991–2001. *Nephrology Dialysis Transplantation*, 17(12):2212–2219.

McFadden, J. (1962). On the lengths of intervals in a stationary point process. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):364–382.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, 38(1):153–168.

Meilă, M. (2007). Comparing clusteringsan information based distance. *Journal of multivariate analysis*, 98(5):873–895.

Molenberghs, G. and Verbeke, G. (2005). *Models for discrete longitudinal data*. Springer, New York.

Murphy, S., Rossini, A., and van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, 92(439):968–976.

Murphy, S. A. and Sen, P. K. (1991). Time-dependent coefficients in a cox-type regression model. *Stochastic Processes and their Applications*, 39(1):153–180.

Nagin, D. S. and Odgers, C. L. (2010). Group-based trajectory modeling in clinical research. *Annual review of clinical psychology*, 6:109–138.

Ning, J., Qin, J., and Shen, Y. (2014). Semiparametric accelerated failure time model for length-biased data with application to dementia study. *Statistica Sinica*, 24(1):313.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, 15(2):780–799.

Parikh, N. and Boyd, S. (2014). Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239.

Parzen, M., Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Troxel, A., and Molenberghs, G. (2007). Pseudo-likelihood methods for the analysis of longitudinal binary data subject to nonignorable non-monotone missingness. *Journal of data science*, 5(1):103–129.

Peng, J. and Müller, H.-G. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. *The Annals of Applied Statistics*, 2(3):1056–1077.

Perlman, R. L., Kiser, M., Finkelstein, F., Eisele, G., Roys, E., Liu, L., Burrows-Hudson, S., Port, F., Messana, J. M., Bailie, G., et al. (2003). Renal research institute symposium: The longitudinal chronic kidney disease study: A prospective cohort study of predialysis renal failure. *Seminars in Dialysis*, 16(6):418–423.

Pollard, D. (2002). *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press, New York.

Qin, J. and Liang, K. (1999). Generalized odds ratio model and pairwise conditional likelihood. Technical report, Technical Report. Department of Biostatistics, Johns Hopkins University, Baltimore, MD.

Qin, J., Ning, J., Liu, H., and Shen, Y. (2011). Maximum likelihood estimations and em algorithms with length-biased data. *Journal of the American Statistical Association*, 106(496):1434–1449.

Qin, J. and Shen, Y. (2010). Statistical methods for analyzing right-censored length-biased data under cox model. *Biometrics*, 66(2):382–392.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

Renard, D., Molenberghs, G., and Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, 44(4):649–667.

Ribatet, M., Cooley, D., and Davison, A. C. (2009). Bayesian inference from composite likelihoods, with an application to spatial extremes. *arXiv preprint arXiv:0911.5357*.

Saran, R., Li, Y., Robinson, B., Ayanian, J., Balkrishnan, R., Bragg-Gresham, J., Chen, J. T., Cope, E., Gipson, D., He, K., et al. (2015). US renal data system 2014 annual data report. *American Journal of Kidney Diseases*, 66(1).

SAS Institute Inc. (2011). *SAS/STAT Software, Version 9.3*. Cary, NC.

Sen, P. K. (1960). On some convergence properties of u-statistics. *Calcutta Statist. Assoc. Bull*, 10(1):18.

Shen, Y., Ning, J., and Qin, J. (2009). Analyzing length-biased data with semiparametric transformation and accelerated failure time models. *Journal of the American Statistical Association*, 104(487):1192–1202.

Shen, Y., Ning, J., and Qin, J. (2016). Nonparametric and semiparametric regression estimation for length-biased survival data. *Lifetime data analysis*, (Accepted).

Sperrin, M. and Buchan, I. (2013). Modelling time to event with observations made at arbitrary times. *Statistics in medicine*, 32(1):99–109.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Su, Y.-R. and Wang, J.-L. (2012). Modeling left-truncated and right-censored survival data with longitudinal covariates. *The Annals of Statistics*, 40(3):1465–1488.

Tan, K. M., Witten, D., et al. (2015). Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2):2324–2347.

Thiébaut, A. and Bénichou, J. (2004). Choice of time-scale in cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in medicine*, 23(24):3803–3820.

Tian, L., Zucker, D., and Wei, L. (2005). On the cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469):172–183.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Topchy, A. P., Jain, A. K., and Punch, W. F. (2004). A mixture model for clustering ensembles. In *SDM*, pages 379–390. SIAM.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge University Press, New York.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence*. Springer.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10(2):616–620.

Vardi, Y. (1989). Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation. *Biometrika*, 76(4):751–761.

Varin, C. (2008). On composite marginal likelihoods. *AStA Advances in Statistical Analysis*, 92(1):1–28.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.

Vonesh, E., Snyder, J., Foley, R., and Collins, A. (2006). Mortality studies comparing peritoneal dialysis and hemodialysis: what do they tell us? *Kidney International*, 70:S3–S11.

Wang, B., Zhang, Y., Sun, W., and Fang, Y. (2016). Sparse convex clustering. *arXiv preprint arXiv:1601.04586*.

Wang, M.-C. (1991). Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86(413):130–143.

Wang, M.-C. (1996). Hazards regression analysis for length-biased data. *Biometrika*, 83(2):343–354.

Wang, M.-C., Brookmeyer, R., and Jewell, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics*, 49(1):1–11.

Wicksell, S. (1925). The corpuscle problem. a mathematical study of a biometric problem. *Biometrika*, 17(1-2):84–99.

Wolfe, R. A., Ashby, V. B., Milford, E. L., Ojo, A. O., Ettenger, R. E., Agodoa, L. Y., Held, P. J., and Port, F. K. (1999). Comparison of mortality in all patients on dialysis, patients on dialysis awaiting transplantation, and recipients of a first cadaveric transplant. *New England Journal of Medicine*, 341(23):1725–1730.

Wu, F., Kim, S., Qin, J., Saran, R., and Li, Y. (2017). A pairwise likelihood augmented estimator for the cox model under left-truncation. *Biometrics*, (In revision).

Xue, L., Zou, H., and Cai, T. (2012). Nonconcave penalized composite conditional likelihood estimation of sparse ising models. *The Annals of Statistics*, 40(3):1403–1429.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika*, 56(3):601–614.

Zeng, D. and Lin, D. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika*, 93(3):627–640.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of statistics*, 38(2):894–942.

Zhao, Y. and Joe, H. (2005). Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, 33(3):335–356.

Zhu, C., Xu, H., Leng, C., and Yan, S. (2014). Convex optimization procedure for clustering: Theoretical revisit. In *Advances in Neural Information Processing Systems*.

Zucker, D. M. and Karr, A. F. (1990). Nonparametric survival analysis with time-dependent covariate effects: a penalized partial likelihood approach. *The Annals of Statistics*, 18(1):329–353.